THE VALIDITY AND RELIABILITY OF THE NINTH GRADE DIRECT WRITING ASSESSMENT: DETERMINING ITS EFFECTIVENESS IN MEASURING ENGLISH LANGUAGE LEARNER STUDENT PROFICIENCY

by

David E. Dunn

A dissertation submitted to the faculty of
The University of Utah
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Teaching and Learning

The University of Utah

May 2011

The University of Utah Graduate School

STATEMENT OF DISSERTATION APPROVAL

The dissertation of **David E. Dunn**

has been approved by the following supervisory committee members:

Doug Hacker, Chair	**12-08-10**	Date Approved
Janice Dole, Member	**12-08-10**	Date Approved
Lauren Liang, Member	**12-08-10**	Date Approved
Randy Honaker, Member	**12-08-10**	Date Approved
John Kircher, Member	**12-08-10**	Date Approved

and by **Elaine Clark**, Chair

of the Department of **Teaching and Learning**

and by Charles A. Wight, Dean of the Graduate School.

ABSTRACT

Many national reports indicate that more attention needs to be placed on writing and the teaching of writing in schools. The purpose of this quantitative study was to, first, examine the structure of the DWA and, second, to use the scores from the DWA to examine the relationship between ELL status and writing proficiency.

Five major research questions were addressed: 1) Does the DWA provide valid and reliable scores of writing proficiency for students in general and for specific groups of students based on ELL status and ethnicity? 2) What is the relation between ELL status and writing proficiency for ninth-grade students attending public schools in Utah during the years 2004, 2005, 2006, and 2007, and to what extent do student variables, such as gender, social economic status, and ethnicity, independently and cumulatively explain the relationship? 3) To what extent do the school variables, percent of low-income students in a school, percent of minority students in a school, size of the school, and mean ELL status at a school independently and cumulatively explain the relation between ELL status and writing proficiency? 4) To what extent do the district variables, percent low-income students in a district, percent minority students in a district, size of the district, mean ELL status in a district, and whether a district is urban or rural independently and cumulatively explain the relation between ELL status and writing

proficiency? 5) To what extent does the relationship between ELL status and writing proficiency interact with ethnicity?

The results of the study indicated that the DWA was a valid and reliable form of writing assessment. Determined also was the fact that deficiencies in writing skills are at their greatest when ethnicity, social economic status, and limited English language proficiency are considered. The results of the present study have suggested ways to rethink how writing is conceptualized and assessed, how past instructional practices have possibly resulted in disparities among ethnic groups, how writing skills vary with student, school, and district characteristics, and how differential writing instruction may benefit students of different ethnic groups and ELL statuses. A "one-size-fits-all" approach to writing instruction will not benefit all students in Utah or throughout the nation. Because writing differs between ethnic groups, writing instruction must differ as well.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

ACKNOWLEDGEMENTS

I would have never been able to complete this journey without the support, encouragement, and patience of my faculty advisor, Dr. Doug Hacker. During the many years of navigating through the challenges of completing a doctoral program, Dr. Hacker has served as a true mentor, expert researcher, teacher, and friend. Thank you for believing in me and always giving me the hope that I would one day be able to complete what I started.

I would also like to thank Dr. Janice Dole, Dr. Lauren Liang, and Dr. Randy Honaker for supporting me and being part of my committee. Their feedback and encouragement helped me complete this part of my journey.

To my parents, Derrick and Heidi Dunn, thank you for teaching me the value of hard work. Your prayers and support helped me to continue to push forward.

To my family, I owe my deepest gratitude for supporting me to the end. I would like to dedicate this work to my wife, Heather, who spent many hours raising our children when I was busy reading, researching, and writing. To my children, Dallin, Mikayla, Collin, Caden, and Landon, who through their many comments and jokes, related to

whether or not Dad would ever finish, helped me keep a true perspective on the important things in life.

CHAPTER 1

INTRODUCTION

Statement of the Problem

Writing remains the "silent R" in the traditional triad of what students need to learn. According to the Nation's Report Card for Writing (National Center for Education Statistics, 2003), a majority of our nation's public school students cannot write at a proficient level. In the 2003 report, *The Neglected "R"* by the National Commission on Writing in America's Schools and Colleges, students in the United States are not at the writing level they need to be (National Commission on Writing, 2003). The results of the 2002 National Assessment of Educational Progress (NAEP) report indicate that more than two-thirds of the students tested in grades four, eight, and twelve were below grade proficiency (National Center for Educational Statistics, 2002). While the most recent NAEP report indicates that writing scores have improved during the past 10 years, 77% of 8^{th}-graders and 86% of 12^{th}-graders are still below the writing proficiency standard (National Center for Educational Statistics, 2007). Moreover, the report has identified that the deficiencies in writing skills are at their greatest when ethnicity, social economic status, and limited English language proficiency are considered. These national reports

clearly indicate that more attention needs to be placed on writing and the teaching of writing.

The inability to write effectively is likely to have an impact on the future employment of students. A survey of 120 human resource directors from several nationwide companies, conducted by the National Commission on Writing (2003), determined that people who are unable to write and communicate effectively stand the chance of not being hired. As educators, we face the challenge of preparing students to use writing as a way to communicate, to learn, persuade others, and as a means for personal self-expression. More and more educators and leaders are realizing that writing is central to success in and out of school (National Writing Project & Nagin, 2006).

In schools, writing is used to share knowledge, learn, and explore feelings and beliefs. It is used as a tool to share results of studies and research. Writers gather and organize their ideas, draft their compositions, revise and edit their drafts, and publish what they have written (Strickland et al., 2001). Because writing is often the primary instrument that teachers use to evaluate academic performance, students' abilities to master writing can influence their success in other academic areas as well (Graham, MacArthur, & Fitzgerald, 1982).

Often, the kinds of writing used in schools are minimal, focusing mostly on spelling tests, lab reports, fill-in-the-blank worksheets and short-answer questions (Graham et al., 2007). Rarely do students write as a means to share knowledge, learn, and explore feelings and beliefs. A survey of high school students' writing experiences indicates that weekly writing took form in responses to literature questions, summaries, expository and essays were a monthly task, while narrative and contrast essays were

written only once or twice a month (Scherff & Piazza, 2005). Even though it is clear that writing plays an important part of what students must do in school, the amount of time that needs to be dedicated to helping students become better writers is unclear. Using the data gathered by the NAEP, it remains unclear how much time should be set aside for writing and writing instruction in order to reach a proficient level as determined by this report. A barrier to improving the teaching of writing within classrooms arises from the fact that researchers currently have incomplete data on how much time students write or what they write (Cutler & Graham, 2008). The reality is that at the elementary level, students spend about 3 hours per week on writing assignments, and secondary students do about one three-page paper a month (National Commission on Writing, 2003). While we may not yet know the amount of writing that should take place in schools, we do know that the amount of writing that does take place is inadequate for students to be successful in an industrialized society (Graham et al., 2007).

Equally relevant for school is the fact that "Writing is not merely a tool for transmitting knowledge; it is also a source of knowledge; it is a resource for dealing with language and thought" (Tolchinsky, 2006, p. 84). Through writing, we learn and record information that helps us understand our nature as well as our history and culture (Freedman & Dyson, 1987). Writing differs from other forms of communication such as speech and reading, in that it leaves visible traces. Because writing leaves traces, it is a very suitable mnemonic device that enables planning, monitoring, revision, and editing in the process of composition (Tolchinsky, 2006). With print on paper, writing makes it possible for students to revisit or reexamine what was expressed. It can be seen as a meaning-making process in which students negotiate meaning with the text produced.

Through writing, students have the ability to convey knowledge and ideas, express themselves, and transmit information. The advent of e-mail, text messaging, and other forms of electronic text have made writing an even more flexible communication tool (Graham et al., 2007).

While writing is an important part of schools and society, it is often a task that many students feel uncomfortable performing. It is an extraordinarily complex activity that incorporates thought processes, feelings, and social interactions (Perin, 2007). We know that when people write, they must draw on a variety of mental operations such as making plans, retrieving ideas from memory, drawing inferences, creating concepts, developing an image of the reader, and testing what they have written against that image (Flower & Hayes, 1980). Other complexities include having to satisfy the constraints of topic, audience, purpose, and of physically creating the text itself, including critical thinking, rhetorical stances, and writing conventions (Bruning & Horn, 2000). Combining all of these processes can be demanding for any writer, including those who write often. Even the most accomplished writers say that writing is challenging, most notably because there is so much uncertainty embedded in the process of doing it (National Writing Project & Nagin, 2006).

Writing is rarely a favorite task for students because of its many challenges. This result is often due to feelings of inadequacy with writing skills and discomfort with the subject matter that is the focus of writing. The difficulties encountered during writing, the lack of feedback from teachers, and the primary focus on writing mechanics rather than conveying meaning tend to discourage students from writing. This discouragement is further exacerbated when students know they will receive a score for their performance.

Part of assessing writing is understanding what constitutes good writing. However, what constitutes good writing is not uniformly agreed upon, even in a single classroom, and when there is agreement, exactly how objective measures of assessment can be fairly assigned is uncertain. Because of the many genres and styles of writing, and the simple fact that writing is not as objectively assessed as other subjects, such as math, science, or computer programming, students are reluctant to engage in writing and often lack interest and motivation to write.

It is not difficult to identify the lack of motivation students experience when it comes to writing. Because of its many complexities, students get frustrated with writing due to a lack of patience, ability, and the understanding of it. Gene Fowler said it best when he described having to write as, to simply sit staring at a blank sheet of paper until the drops of blood form on your forehead. Bruning et al. (2000) identified that the lack of motivation to write is due to the set of beliefs about writing, the purpose of it, and its complexity. In contrast, when students feel confident about their writing abilities, their performance improves. According to Pajares (2003), students' confidence in their writing capabilities has an influence on writing motivation as well as the writing outcomes in school. These outcomes are measured in the students' abilities to have proper skills: grammar, usage, composition, and mechanical writing skills, as well as to complete a task, organize sentences into paragraphs and develop clearly expressed ideas.

As challenging as writing may be, teaching it is even more difficult. Writing is complex, and so is the amount of instruction required to help students reach the high standards of learning expected of them (National Writing Project & Nagin, 2006). In many cases, the difficulty in teaching writing is that few tasks involve so many complex,

interwoven layers. Writing demands an understanding of the content, knowledge of the audience and the context, and the ability to use appropriate conventions for that audience and context.

An even greater challenge lies in teaching writing to the many students who live in this country and whose native language is not English. Writing in one's native language (L1) is a demanding task that calls upon several language abilities, as well as upon more general metacognitive abilities. Writing in a second language (L2) is even more demanding because several of these constituent abilities may be less well-developed than in one's L1 language (Schoonen et al., 2003). While an L1 writer may not fear the lack of English knowledge, writing can be difficult because it is a struggle of thought and imagination to find an expression clear enough to put into words. The deficit L2 writers possess starts with a lack of knowledge about the English language.

Developing proficiency in writing presents a special difficulty for English Language Learner (ELL) students, many of whom believe they cannot write in their first language, much less a second language (Kasper & Petrello, 1997). The "I can't write English syndrome" (Thomas, 1993) leads to anxiety and thereby inhibits progress in achieving writing proficiency. Because of insecurity due to a lack of English and English writing abilities, the focus of writing becomes more syntactic. Breetvelt, van den Bergh, and Rijlaarsdam (1994) point to the fact that less-skilled writers — in this case, L2 writers — concentrate on corrections pertaining to punctuation, syntax, and the meaning of individual words, while L1 writers tend to diagnose problems that have bearing on the meaning of the text as a whole so that their revision more frequently applies to

paragraphs or larger sections of text. In other words, language proficiency has a big impact on writing outcomes.

Through my own experience as a school administrator, I have had many conversations with other school principals regarding writing. Despite all the national calls to improve writing at all grades, it has become clear that more needs to be done in schools regarding writing and writing instruction — especially in addressing the needs of ELL students. Writing remains something that is done primarily in language arts classes, and the instruction of it continues to focus on grammar at the expense of meaning. Although Utah does have a standard writing curriculum that all students must take, it is ill-defined.

Even more alarming is the fact that the increase in ELL students has dramatically accelerated over the past 20 years, and yet little has been done to meet their writing needs. According to the Bureau of Economic and Business Research (2003), the state of Utah is one of the fastest-changing states in the nation in terms of diversity; the minority population is expected to increase 278% by 2025. This dramatic change in the student population brings with it the need for educators to understand and apply through their teaching the current educational research that can support student learning. Unfortunately, very little is known about the academic needs of ELL students, and even less is known about their literacy needs. According to the recent National Assessment of Educational Progress (National Center for Educational Statistics, 2007), the percent of proficient and advanced eighth-grade writers in Utah has increased by about 8% to 31% from 2002; however, the percent of basic and below-basic writers remains alarmingly high, 69% of eighth-grade students tested below the proficiency level. How ELL students

perform is not reported, and statistics on students by ethnicity is incomplete. Results for Hispanic students indicate that nearly 90% are at basic or below basic proficiency. If we can assume that a large percentage of this group is ELL, then we are in a truly alarming situation.

In addition to the NAEP, which is conducted every 4 years at the national level, when it comes to writing assessments within our own state, the Utah State Office of Education (USOE) only requires sixth- and ninth-grade students to take the Direct Writing Assessment (DWA). Despite the fact that these large state-wide assessments are conducted regularly, the results of these assessments are rarely used to provide feedback to administrators or teachers as to differences among student groups and what these differences mean in terms of instruction. Before any reasoned action can be taken and policy changes made, more needs to be known about the writing skills of students, in particular ELL students, and the variables that may be helping or hindering their progress to writing proficiency.

Research Questions

The purpose of this study was to examine the writing skills of students, particularly ELL students, in the state of Utah and identify variables that may contribute to those skills. This study used archival data provided by the DWA of ninth-grade students who attended Utah's public schools during the 2004, 2005, 2006, and 2007 school years. This sample represents approximately 135,000 students. My initial goal was to examine the validity and reliability of the DWA scores for all students and for subgroups that were determined by ELL status and ethnicity. The DWA must be

providing valid scores that assess the psychological construct of writing skill, and those scores must be measured with acceptable precision. If either validity or reliability is not acceptable, for all students or the identified subgroups, then information provided by the DWA is not useful.

With validity and reliability established, I assessed the relative contributions of student, school, and school district to the relationship between ELL status and writing proficiency. In addition, I identified predictor variables at the individual, school, and district level that might contribute to the relationship. Some of the predictor variables chosen for this study are variables that have been linked to writing proficiency, such as students' gender, income, and ethnicity. By identifying variables that significantly contribute to the relation between ELL status and writing proficiency, I was able to identify the conditions that contributed to or detracted from ELL students' writing proficiency. Once these conditions were identified, the groundwork was established for future empirical work that could more specifically investigate the school and classroom environments that promote writing skills. For example, my analyses show that ELL students thrive in smaller schools in which there is greater diversity. Examining the writing practices in those schools could do much to inform curriculum policy decisions. Accordingly, my research questions were:

1) Does the DWA provide valid and reliable scores of writing proficiency for students in general and for specific groups of students based on ELL status and ethnicity?
2) What is the relation between ELL status and writing proficiency for ninth-grade students attending public schools in Utah during the years 2004, 2005,

2006, and 2007, and to what extent do student variables, gender, social economic status, and ethnicity, independently and cumulatively explain the relation?

3) To what extent do the school variables, percent low-income in a school, percent minority in a school, size of the school, and mean ELL status at a school independently and cumulatively explain the relation between ELL status and writing proficiency?

4) To what extent do the district variables, percent low-income in a district, percent minority students in a district, size of the district, mean ELL status in a district, and whether a district is urban or rural, independently and cumulatively explain the relation between ELL status and writing proficiency?

5) To what extent does the relationship between ELL status and writing proficiency interact with ethnicity?

CHAPTER 2

REVIEW OF LITERATURE

Over the past 2 decades, the study of writing has increasingly become part of the mainstream in applied linguistics. Connor (1996) said, “The reasons for this change are many: the increased understanding of language learners’ needs to read and write in the target language; the enhanced interdisciplinary approach to studying second language acquisition through educational, rhetorical, and anthropological methods; and new trends in linguistics” (p. 5). With open admission policies and diverse populations of students seeking different levels of education, and the rapidly increasing demands of skilled writers in the job market, writing has become a major area of interest. While writing in the primary language that one speaks is not an easy process, learning to write in a second language can be a demanding task. Second-language writing researchers such as Silva (2001) have documented the difficulties of L2 writing and that no single approach exists to solve the problem (Blanton & Kroll, 2002). From current research in the field of writing, specifically cross-cultural aspects of second-language writing, one of the main factors that influence the variation between the writing of first-language (L1) and second-

language (L2) writers is culture. However, preliminary to addressing how culture affects writing, I described contrastive rhetoric with the goal of highlighting differences in writing between L1 and L2 writers. I then focused on the challenges of using current writing assessments to assess the writing of L1 and L2 writers.

Contrastive Rhetoric

Before showing how L1 and L2 students differ in their writing, it is important to describe contrastive rhetoric and its history. The origins of contrastive rhetoric are rooted in the Sapir-Whorf hypothesis (Matsuda, 2001). The Sapir-Whorf hypothesis suggests that different languages affect perception and thought in different ways. The "strong" version of the hypothesis suggests that a person's native language determines his or her thought and perception; however, the "weak" version maintains that a person's native language only influences his or her thought and perception (Connor, 1996; Martin, 1992). Each of these lines of thought has been studied with varying results (Matsuda & Silva, 2005). Although support for the strong version has waned considerably in recent years, the weak version still has credibility. Robert Kaplan, whose pioneering work has had considerable impact on the study of L2 learners, was mainly influenced by the weak version of the Sapir-Whorf hypothesis (Connor, 1996; Matsuda, 2001).

The term "contrastive rhetoric" was coined in 1966 by Kaplan, who, along with other writing instructors, discovered that the writing patterns of international students who had recently come to the United States were much different from the writing patterns of native English writers (Panetta, 2001). Kaplan started his research by looking at the writing of English as Second Language (ESL) students to determine how their writing

deviated from that of native English language writers. Through his analysis on writing by ESL students, he noted that the differences were not simply grammatical or surface matters, such as spelling, but instead the differences were paragraph order and structure (Kaplan, 1988). From this came the observation that instead of sentence-level errors being the focus of writing, the larger written product became the area of research. Kaplan's article "Culture Thought Patterns in Intercultural Education" was the first in the field of ESL that focused on the rhetoric of writing and extended analysis beyond the sentence level.

Kaplan's work continued when he compared ESL cultural practices to typical Western practices and discovered many interesting rhetorical trends and deviations (Piper, 1985). Kaplan analyzed 600 ESL texts written by students from Semitic, Oriental, Romance and Russian language backgrounds and grouped them into five categories according to patterns related to their L1 background. He discovered that students from Anglo-European languages preferred writing in a linear style, whereas students from Asian languages seemed to take a more indirect approach by making their observations at the end of the paper (Panetta, 2001). The linear style is the development of writing that follows a "direct" series of outlined steps. It begins with an initial paragraph that contains a thesis statement supported by additional paragraphs. The paragraph development in writing done by students from Semitic languages tended to be based on a series of parallel organizations of coordinate, rather than subordinate clauses, whereas students from Romance and Russian languages tended to prefer extraneous material (Connor, 1996). In the parallel organization, ideas and sentences are much less indirect. Words such as "and" are used extensively to connect ideas. These discoveries helped Kaplan to

note that rhetorical structure was culture-dependent. Connor (1996), who focused on Kaplan's work, referred to contrastive rhetoric as "an area of research in second language acquisition that identifies problems in composition encountered by second language writers and, by referring to rhetorical strategies of the first language, attempts to explain them" (p. 5).

Since the work done by Kaplan, there have been both proponents of his work, as well as opposition to it. Leki (1992), a proponent, pointed out that "Even though writing instructors who teach ESL students may not have backgrounds in the rhetoric of different cultures, contrastive rhetoric helps bypass stereotypes and realize that writing strategies are culturally formed" (p. 138). Purves (1988) pointed out that,

> When students, taught to write in one culture, enter another and do not write as do the members of the second culture, they should not be thought stupid or lacking in 'higher mental processes,' as some composition teachers have stated, instead, they simply do not know about the rhetorical structures of the new culture, but they have the capability to learn the new conventions if given ample opportunity. (p. 19)

In short, a number of researchers have argued that, with contrastive rhetoric, instructors who teach writing to ESL students can come to see that our truth is not necessarily the objective truth and that, in reality, truth is a relative concept across cultures and languages (Leki, 1991).

Opponents to Kaplan's work have had concerns about some of the research on early contrastive rhetoric. These criticisms focus mainly on the fact that he lumped many different languages together. An example of this is the grouping of all Asians as "Orientals." Moreover, in addressing the writing patterns of L1 and L2, he oversimplified their forms of writing, particularly when it came to attempting to see L1 thought patterns in L2 essays (Hyland, 2003).

Martin (1992) criticized Kaplan's work for being too simplistic in its research methodology and conceptualization. Leki (1991) stated that contrastive rhetoric was too ethnocentric by privileging the writing of native English speakers and for characterizing English rhetoric as linear and direct. Grabe and Kaplan (1996) pointed out that the difficulties of contrastive rhetoric research concentrated on the final product without giving adequate attention to ways in which the text was produced. Severino (1993) suggested that it was not possible to detect cultural thought patterns at the paragraph level, especially if the paragraph was revised.

According to Connor (1996), "Kaplan's first study of contrastive rhetoric provided a model of writing for a theory of second language teaching that is more useful in some applications than in others" (p. 8). Kaplan's model, for example, is not particularly relevant for the theory of translation, because it refers to second-language texts only when speculating about first-language influence. A model for translation needs to compare text in both languages: the source language (in which the text was originally written) and the target language (into which the text was translated). While Kaplan's model is indeed useful in evaluating second language written products, a different model is needed to describe differences in composing processes across cultures (Connor, 1996).

Culture

Although Kaplan's work focused on the transfer of first-language cultural conventions to second-language performance, he did not pay much attention to the reasons for culture-specific writing styles. Studies have shown the effects that culture has

on writing. Ward Goodenough (1964), a cultural anthropologist and linguist, gives the following definition of a culture:

> By definition, we should note that culture is not a material phenomenon; it does not consist of things, people, behavior, or emotions. It is rather an organization of these things. It is the forms of things that people have in mind, their models for perceiving, relating, and other wise interpreting them. As such, the things people say and do, their social arrangement and events, are products or by-products of their culture as they apply it to the task of perceiving and dealing with their circumstances. (p. 36)

From this definition, a better sense can be gained of how culture impacts writing. According to Panetta (2001), "Language, as we know, differs among nations and geographic locations, but differences even within the same language are rooted in the cultural history of the native speaker" (p. 48). For example, the Spanish spoken in Mexico is quite different from the Spanish spoken in Spain; English spoken in Britain differs from the English spoken in America. Further, within the United States, dialects differ among regions and states. For example, a Boston dialect makes the pronunciation of park and car into *pahk* and *cah*, and a New Orleans dialect makes them into *pawk* and *caw*.

Understanding the relationship of culture, language, and rhetoric is confusing enough within the confines of one culture, but negotiating these relationships across multiple cultures can be challenging in the field of writing (Panetta, 2001). Meeting the rhetorical expectations of another culture requires more than attention to language. According to Hoft (1995), educational variables such as literacy, learning styles, and common body of knowledge affect the way people understand and respond to written material. One's own culture and beliefs frame the manner in which ideas are organized and expressed through writing.

Before looking at some of the studies that identify differences in writing among cultures, it is important that we understand some of the origins of our own language and culture of writing and reading. The English language and its related thought patterns have evolved out of the Anglo-European culture. Kaplan (2001) points out that "The expected sequence of thought in English is essentially a Platonic-Aristotelian sequence, descended from the philosophers of ancient Greece and shaped subsequently by Roman, Medieval European, and later Western thinkers" (p.12). As speakers and writers of English, we expect as an integral part of communication a sequence that is dominantly linear in its development. An English expository paragraph usually begins with a topic statement followed by a series of subdivisions of that topic statement. Each subdivision is supported by examples and illustrations that (a) develop the topic statement, (b) relate it to all other ideas in the whole essay, and (c) employ it in proper relationship with other ideas that prove or argue some position. Just as the English language has its own style, other cultures through history have developed their own styles of writing. Several illustrations of the differences in writing based on culture are presented next.

In English, subordination is preferred in many situations and is taught to students through sentence-combining and other syntactic methods, whereas in Arabic, coordination is preferred. For example, in English, we would say the following: *The boy was here, and he drank the milk*. In Arabic, the sentence would be written as: *Milk was drunk by the boy who was here* (Kaplan, 1972).

William Eggington (1987) showed that Korean texts are characterized by indirectness and nonlinear development. Korean text starts with an introduction, the development of a topic, a turning to a somewhat unrelated topic, and then a conclusion.

Korean and English are similar in some aspects; however, inserting an "unrelated topic" as part of an essay might cause an English reader to misunderstand the full meaning of the essay.

Studies on Chinese writing by Scollon (1991) and Japanese writing by Hinds (1990) share a common theme when compared to English: Both languages have a delayed introduction of purpose. Scollon attributes the indirectness in Chinese writing to a different view of self in Chinese culture from the Western image of selfness. Because of the concept of self, Chinese writers find it difficult to be direct and to express a personal point of view in a thesis statement at the beginning of a piece of writing. Hinds (1990) found that Japanese students tend to place the main idea at the end of paragraphs. Kubota (1992) pointed out that Japanese students' perception about the differences between English and Japanese are that:

> Japanese text is indirect, ambiguous, roundabout, illogical, digressive, has the main idea at the end and contains a long introductory remark and long, complex sentences; English is direct, clear, logical, has the main idea stated at the beginning, and has unity in the paragraph and little digression. (p. 44)

While relatively few studies have examined German-English contrasts, Clyne (1987) indicates that English writers make their texts as readable as possible, whereas German writers emphasize content. English writing, through its linear process, provides the information needed for the reader to understand. German, in contrast, is written in such a way that it is the reader's responsibility to try and make connections.

Montano-Harmon (1988) and Lux and Grabe (1991) examined a number of Spanish-English rhetorical contrasts. They found that Spanish writers prefer a more elaborate style of writing. Spanish writers use more adjectives and causal conjunctions, while English writers use simple vocabulary, few synonyms, and much shorter sentences.

These studies identify that there are differences between L1 and L2 writers. This research helps support the idea that L2 students bring with them a writing culture that can be different in many ways from the standards of written English within this country. Hyland (2003), in a summary of the collective works of Connor (1996), Grabe and Kaplan (1996), and Hinkel (1999), describes differences between L1 and L2 student academic essays in the following ways:

- different organizational preferences
- different approaches to argument (justification, persuasive appeals, credibility)
- different ways of incorporating material (use of quotes, paraphrase, allusion, unacknowledged borrowing, etc.)
- different ways and extent of getting reader's attention and orienting them to topic
- different estimates of reader knowledge
- different uses of cohesion and meta-discourse markers
- differences in how overt linguistic feathers are used (generally less subordination, passives, modifiers, lexical variety, and specificity in L2 writing)
- differences in objectivity (L2 texts often contain more generalizations and personal opinions)
- differences in complexity of style

While this list identifies some of the differences that exist between L1 and L2 writers, we need to understand that not all L2 writers share the same difficulties when writing. Understanding the culture of L2 writers can provide insights into the areas of writing that may be potentially problematic. Alternatively, understanding the culture of

L2 writers also can provide insights into new conceptualizations of the world through their unique approaches to writing.

Given that differences in writing do exist due to culture, applying a single uniform approach to writing assessment may not be appropriate. Methods of assessment that are used in L1 writing may need to differ from those that are used in L2 writing. For example, assessing the organization of an essay written by a native Chinese or Japanese speaker on the basis of accepted English construction could result in significantly lower scores. The writer may have strong writing skills but lack only knowledge of the conventions of written English. Interpreting the low scores as a lack of writing skill would be a mischaracterization of the writer's ability. Further, instruction for this person should be directed at writing conventions and not at basic writing skills.

Writing Assessments

The assessment of writing is perhaps one of the most controversial topics related to educational assessment in this country (Yancey, 1999). Given the potential impact that culture has on writing, questions about the assessment of writing have not been easily resolved. While the primary purposes of many assessments are used to compare individual achievement to an established norm, the assessing of writing brings with it other challenges.

For more than a century, writing assessment discussions have focused on whether to use direct or indirect assessments. Within the past 60 years, the debate has intensified. Yancey (1999) identified "three waves" of writing assessment: objective tests, holistic scoring, and portfolios. The first wave occurred from 1950 to 1970, when the focus was

objective tests. Writing was assessed indirectly using writers' responses to objective multiple choice questions concerning a variety of writing components. Such indirect writing assessments assess writing ability by testing a subset of skills assumed to be integral to writing ability (Grabe & Kaplan, 1996). Testing is often done using multiple-choice measures that test students' grammar, vocabulary, and written expression. This kind of assessment focuses writers more on the details of grammar so that mistakes are not made, rather than on idea development or paragraph cohesion. Once considered the primary measure for writing, the use of indirect writing assessments has decreased in the past 10 years.

From 1970 to 1986, the second wave of writing assessment turned to holistic scoring, which was used for both L1 and L2 writers (Weigle, 2002). In contrast to indirect writing assessments, this method provided a more accurate assessment of actual writing performance. Holistic scoring involves assigning each essay a single score based on its overall quality. It is defined as a type of assessment scoring in which scorers provide one overall score based on their overall impression of the quality of writing, as opposed to a variety of scores that quantify strengths and weaknesses (Baldwin, 2004). It was the most common form of scoring for large assessments. Rather than giving points throughout the paper, the scorer reads the text and assigns a single numerical score for the paper. For example, a teacher might require students to write an essay on photosynthesis. Rather than giving points for using specific vocabulary or describing the steps a plant goes through to make food, the scorer can give an overall numerical score to the paper. Variations of this method of writing assessment continue to this day.

The third wave of writing assessment, from 1986 to the present, which developed primarily in response to criticisms of timed essay examinations, has focused on the use of writing portfolios. A portfolio consists of a collection of a student's writing that is gathered over a period of time and used to demonstrate developments in writing ability. Supporters of portfolios believe that students have the opportunity to revise and develop their writing rather than have their writing assessed using a single timed essay (Elbow & Belanoff, 1991).

In our American culture, writing assessments are developed with an English-speaking student in mind and with the cultural contingencies associated with an English-speaking culture. The writing assessments are created with the intent of having the writer write in a dominantly linear style with a topic sentence in the first paragraph, followed by additional paragraphs that develop and support the primary idea, and ending with some sort of a summary. However, this "one-size-fits-all" method of assessment may not meet the needs of all students. For students whose first languages differ from English, this form of writing may be alien to them and difficult to achieve. Therefore, using a method of assessment that is based on English conventions may be more a measure of how well the student has conformed to writing in English and less a measure of a student's writing proficiency. The student may be quite proficient in his or her native language.

Knowing how to fairly and validly assess writers whose first language is not English can be a challenge. According to Silva (2001), "There exists at present, no coherent, comprehensive theory of L2 writing" (p. 201), and without a comprehensive theory of L2 writing, methods of assessment are also lacking. A reason for this deficit in comprehensive theory can be explained in part by the newness of L2 writing as an area of

inquiry. Another equally important reason is the prevalent assumption that L1 and L2 writing are, for all intents and purposes, the same (Silva, 2001) and therefore, should be assessed the same. This unexamined assumption has led L2 writing specialists to rely for direction almost exclusively on L1 composition theories, theories that are, incidentally, largely monolingual, monocultural, ethnocentric, and fixated on the writing of English-speaking undergraduates in North American colleges and universities.

Because L2 writing likely differs from L1 writing, using a single assessment system that has been devised to measure the writing of L1 writers might put L2 writers at a considerable disadvantage. According to Garcia and Pearson (1994), English language learners may face cultural as well as linguistic biases when taking standardized tests. Factors that may affect an L2 student's writing assessment include testing anxiety, vocabulary, time constraints, and simply understanding what the writing prompt may be asking. There are writing genres in English that may not exist in other cultures, and if there are genres common across cultures, how they are structured across those cultures may differ. The persuasive genre, which is often used in writing assessment, takes different forms in different cultures. In English, the persuasive essay takes the form of logical arguments that are to appeal to the reader's sense of reasoning. In Chinese or Japanese, a persuasive essay is based more on eliciting an emotional response rather than a logical one (Hinds, 1990; Scollon, 1991). Simply stated, an L2 student taking a writing test in this country may be taking a biased test, due to the fact that it was developed and normed for the dominant culture. Alexander and Parson (1991) made this clear when they stated:

> Educational testing and school assessment, like the students for whom they are targeted and the teachers who are often their creators/administrators, do not have

> a life only inside the schoolhouse or during school hours. Students and teachers live in a social and cultural milieu that includes the school, the home, and the community as well. Educational testing and school assessment are elements of the same society and culture that contain students, their families, and their teachers. Thus, tests and assessments are subject to the same social and cultural influences as other parts of the social realm. Because of this influence, any discourse on the issues of educational testing and school assessment must take into account social and cultural factors. (p. 244)

An assessment system that emphasizes clear, logical, and direct development of argumentation will discredit the writing from cultures that emphasize indirectness and ambiguity. Which characteristics of writing are measured by an assessment system and how validly those characteristics are measured across different cultures are concerns that have not been adequately addressed. With the tremendous increase in multicultural classrooms, these concerns become increasingly important.

Writing ability has received much more attention in recent years than it had previously, both with respect to research, employment, and access to higher education. For example, standardized tests such as the SAT, ACT, and GRE now all include writing components. Teachers' responses to student writing can greatly influence the students' attitudes about writing, and providing them with fair and supportable assessments can have a great positive impact on their motivation to write. While unclear, vague, or ambiguous feedback can lead to frustration, supportive and constructive feedback can lead to higher motivation and creativity. To provide supportive and constructive feedback, a writing assessment cannot fault students on the basis of cultural differences. Rather, the assessment must clearly focus on the student's ability to produce meaning using a consensual writing system.

Grabe and Kaplan (1996) pointed out that, "Until the emergence of the process movement in writing instruction, much feedback to students on their writing appeared in

the form of a final grade on a paper, often accompanied by much red ink throughout the essay" (p.378). The rationale behind this way of grading was that the student would learn from their mistakes and not replicate them in the future. However, it has become apparent to both teachers and researchers that this form of assessment has often left students feeling defeated and unmotivated to write. The writing-process approach, on the other hand, has had a more positive impact on students because it allows for the student to write through a process (prewriting, first draft, revising, and final draft), which in turn allows for a teacher to give ongoing feedback as the work develops, rather than one assessment on the entire final product.

The type of writing assessment used with student writing varies with the reasons for writing. Teachers might give a writing assessment to measure students' content knowledge gained on an assigned unit. In some cases, the purpose might be for enrollment at a college or university. Another reason may be to determine the proficiency writing level of a student, which is sometimes used to place the student in a particular class or program. In still other cases, national assessments such as the NAEP and state assessments such as the one used in the state of Utah, the Direct Writing Assessment (DWA), are used to gauge and compare the writing proficiency of different groups of students. Questions the present research addresses is whether the DWA is providing valid and reliable scores for all students, but for ELL students in particular, and if so, whether there are factors that can be identified that either contribute to or detract from ELL students' writing proficiency.

The Utah Direct Writing Assessment (DWA)

Different types of scoring rubrics can be used, such as holistic scoring, primary trait scoring, and analytical scoring. As mentioned earlier, holistic scoring involves assigning a single score to an essay based on its overall quality. Often, scores are assigned on the basis of the reader's first impression of the essay, as opposed to a variety of scores that quantify strengths and weaknesses (Baldwin, 2004). Holistic scoring was once the most common form of scoring for large assessments.

Primary trait scoring focuses on particular characteristics of a piece of writing. According to Spandel and Stiggins (1990), "the scoring is based on the premise that all writing is done for an audience and that successful writing will have the desired effect on the audience, mainly due to the impact of the primary, or most important trait within that piece of writing" (p. 5). With this type of scoring, writing assignments are designed to reflect a specific task that a student must carry out. An example of a large-scale assessment that uses this form of scoring is the NAEP test. In it, students are required to complete a task, such as writing a persuasive essay or a letter. The score is based on how well the task was completed.

The large-scale writing assessment that is used by the NAEP uses this form of scoring. In it, students are required to complete a task, such as writing a persuasive essay. Essays are assessed on a 0-to-300 scaled score, with five increasingly proficient levels of writing identified within the scale (i.e., insufficient, uneven, sufficient, skillful, and excellent). Although a single score is assigned to an essay, there are specific writing criteria that are being assessed. These criteria measure the focus and organization of the essay, an identified, consistently supported position, transitions, varied sentence structure,

varied and interesting word choice, and the minimization of spelling and grammatical errors that could interfere with the reader's understanding.

Similarly, the analytical writing assessment of the Graduate Record Examination (GRE) uses primary-trait scoring. Essays are assessed by being assigned a single value on a six-point scale, with higher scores indicting greater proficiency. Once again, although a single score is assigned to an essay, specific criteria have been identified that are assessed. These criteria include analysis of complex ideas, support of main ideas, persuasive examples, focus and organization, varied sentences, precise vocabulary, sentence structure, language usage, and minor spelling and grammatical errors that could interfere with the reader's understanding.

The DWA uses analytical scoring, which consists of isolating several features of writing and assigning each a separate score. Its purpose is to attempt to define characteristics of writing such as conventions and sentence fluency (Spandel & Stiggins, 1990). This form of scoring allows a student to receive a grade and feedback on specific aspects of writing. The scorer can give points for specific characteristics within the writing. Organization, clarity, description, and process are parts that are often included in a single score, but they can be scored separately.

The DWA uses six features for its scoring: (a) ideas, (b) organization, (c) voice, (d) word choice, (e) sentence fluency, and (f) conventions. These same features, known as the Six Traits of Writing, are used in the Utah Core Curriculum, and they are embedded in writing instruction from grades three to 12. Utah students are regularly assessed using the Six Traits of Writing, which are considered by many to be the six most

salient qualities that are inherent in good writing. Currently, the Six Traits are used in school systems in 38 states across the United States.

With the DWA, students are given a prompt, often to write a persuasive essay, to which they must respond within a set amount of time. Once completed, all writing samples are submitted to the Utah State Office of Education (USOE). The USOE has contracted with Pearson Publishing Company to score the DWA for grades six and nine. Each writing sample is scored for each trait on a 1-to-6-point scale, and a composite score is calculated. Although Pearson has provided a thorough description of the scoring procedures of the DWA and provides reports of inter-rater reliability, validity and reliability of the DWA scores is not well-reported.

Research Questions

Research over the past 40 years overwhelmingly supports the fact that there are differences between L1 and L2 writers and that these difference are linked to cultural differences in the approach to writing. A purpose of this study was to examine whether differences between L1 and L2 writers are reflected in the writing scores of the DWA. A second purpose was to examine whether differences between L1 and L2 writers vary by person variables, such as gender and ethnicity; school variables, such as size and diversity; and district variables, such as rural/urban and social economic status. Finding differences along any of these variables helps to inform future research that can specifically investigate personal or environmental factors that either contribute to or detract from the writing skills of ELL students. The ultimate goal of the present research was to stimulate and direct follow-up research that can identify potential solutions to

supporting and helping L2 students become more proficient in the area of writing. These solutions may involve curricular changes at the school, district, or state levels.

Because validity and reliability information on the DWA are not adequately reported, a third purpose of the present research was to investigate both the validity and reliability of the DWA. The DWA must be providing valid scores that assess the psychological construct of writing skill, and those scores must be measured with acceptable precision. This is true for all students and particularly for ELL subgroups. If validity or reliability is not acceptable, then information provided by the DWA is not useful. Therefore, the first research question was:

1) Does the DWA provide valid and reliable scores of writing proficiency for students in general and for specific groups of students based on ELL status and ethnicity?

Once validity and reliability were established, I assessed the relative contributions of student, school, and school district to the relationship between ELL status and writing proficiency. Moreover, predictor variables at the individual, school, and district level that might contribute to the relationship were examined for their potential contributions to the relationship between ELL status and writing proficiency. Therefore, the next three research questions were:

2) What is the relation between ELL status and writing proficiency for ninth-grade students attending public schools in Utah during the years 2004, 2005, 2006, and 2007, and to what extent do student variables, gender, social economic status, and ethnicity independently and cumulatively explain the relation.

3) To what extent do the school variables, percent low-income students in a school, percent minority students in a school, size of the school, and mean ELL status at a school independently and cumulatively explain the relation between ELL status and writing proficiency?

4) To what extent do the district variables, percent low-income in a district, percent minority in a district, size of the district, mean ELL status in a district, and whether a district is urban or rural, independently and cumulatively explain the relation between ELL status and writing proficiency?

Finally, because of the cultural differences that belie ELL status, any relationship between ELL status and writing proficiency may vary by ethnicity. For example, students from an Asian country, in which cultural differences from English-speaking Americans may be quite pronounced, may have much greater difficulty acquiring the conventions of English writing or the organization of writing than students from Central American countries, in which there are still cultural differences from English-speaking Americans, but these differences may not be as pronounced as in the former case. Each of the six traits may vary not only by ELL status but by cultural differences that are represented by ethnicity. Therefore, the last research question was:

5) To what extent do each of the six traits vary with ELL status and ethnicity, and are there interactions between ELL status and ethnicity on any of the six traits?

CHAPTER 3

METHODS

The present study was a quantitative study that had two general purposes. By pursuing these two purposes, I was able to answer my five research questions. The first purpose was to examine the structure of the DWA. By examining various elements of the structure of the DWA and making a judgment whether each element meets specific technical criteria, a body of evidence was produced that can be used to support the use of the DWA scores in making valid and reliable decisions about individual and group writing proficiency. The second purpose was to use the scores from the DWA for ninth-grade students who attended Utah's public schools during the 2004, 2005, 2006, and 2007 school years to examine the relationship between ELL status and writing proficiency.

Structure of the DWA

I examined the structure of the DWA to answer research question 1:

1) Does the DWA provide valid and reliable scores of writing proficiency for students in general and for specific groups of students based on ELL status and ethnicity?

I followed five guidelines for examining whole tests that are described by Osterlind (2006). First, I reviewed the published information about the DWA that is provided by the test's author. This information included descriptions of: (a) the constructs used to evaluate writing (i.e., the six traits), (b) how widely used and accepted these constructs are in the evaluation of writing, (c) how the constructs are measured, (d) what the procedures are for scoring essays, and (e) how inter-rater is maintained at high levels.

Second, for each of the years that I have DWA data, I calculated summary statistics for each of the six traits. These statistics included mean, mode, median, standard deviation, variance, range, skewness and kurtosis. In addition, these summary statistics were calculated for the targeted groups of this study (i.e., ELL status by ethnicity). These statistics provided essential information concerning the distributions of the scores and suggest possible avenues for further statistical analyses. The statistics also allowed me to examine the stability of each of the traits for all students and for students in the targeted groups across several years.

Third, summary statistics from the DWA were compared to similar summary statistics from the National Assessment for Education Progress (NAEP) for Writing 2007. The NAEP is a large-scale writing test that is administered nationally every 4 years to students in eighth and 12^{th} grade. For my comparisons, I used only the eighth-grade data.

I looked for similar trends in the data across the two tests that included gender, ethnicity, and socioeconomic status.

Fourth, I examined information about the reliability of the scores from the DWA. Because reliability is a necessary but insufficient condition for validity, I first examined reliability and then validity. Reliability refers to the precision in psychological measurement. It is an estimate of how well samples of observed scores on a test reflect true scores (i.e., how accurate are the scores in a test). Reliability provides a measure of the magnitude of error in psychological measurement; in this case, the amount of error in the scores from the DWA.

In classical test theory, reliability of a test is defined as the squared correlation between observed-score and true-score variance. However, true scores are never known and can only be estimated; therefore, reliability is often expressed as a correlation between two observed scores. I reported reliability coefficients, Cronbach's α, as an estimate of the reliability of the DWA scores. Because reliability information is not reported by the publishers of the DWA, this information is particularly relevant to using the DWA scores in the current research. In addition, because I was interested in the use of the DWA with ELL students, I calculated Cronbach's α for ELL students and compared the precision in DWA scores for this group with the entire sample.

Fifth, I next turned my attention to the validity of the scores from the DWA. "Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests," (American Educational Research Association, & National Council on Measurement in Education, 1999, p. 9). I took a four-pronged approach to evaluate the validity of the DWA scores. In this approach, I

collected various kinds of empirical evidence and theoretical rationales to evaluate the proposed interpretations of the test scores in light of the purpose of the DWA. The evidence and rationales were organized as construct, content, external, and internal validity, and these various sources of evidence and rationales were examined in a unifying way so that any decisions based on the DWA scores could be supported.

The first prong of my examination of validity concerned the underlying construct that is measured by the DWA. That construct is writing skill, which is defined by six characteristics or traits: voice, organization, ideas, conventions, word choice, and sentence fluency. To establish some support for the construct validity of the six traits, I described the intended meaning of each and examined other writing assessments for evidence of these same traits.

The second prong of my examination of validity was to examine the extent to which the six traits are used in instruction and assessment in schools across the United States. The ultimate goal was to demonstrate that the six traits are widely recognized as defining characteristics of writing proficiency, and hence, are widely used in instruction and assessments.

The third prong of my examination was to analyze the internal structure of the DWA. Using 1 year of data, I first conducted an exploratory factor analysis to investigate the dimensionality of the test. In many writing assessments, there are often two dimensions identified: one dealing with the mechanics or conventions of writing and one dealing with the content and how that content is expressed. The factor analysis uncovered whether this dimensionality exists for the DWA. This analysis was followed by a confirmatory factor analysis using structural equation modeling (SEM). Questions

concerning whether the identified model of the test's internal structure is the best fit or if alternative models can provide a better fit was answered by this analysis. Along with the structural equation modeling, I examined whether the same identified dimensionality of the DWA was present for the targeted subgroups of interest in this study. If cultural or linguistic differences influence writing, then alternative models of the internal structure of the DWA may better describe the data.

The fourth prong of my examination of validity was to analyze data from other years, using the methods described. If the results from the factor analysis and confirmatory factor analysis are replicated for each year, then stronger statements about the generalizability of the test can be made. The degree to which evidence of validity based on one situation can generalize to other situations is reflective of the test's overall validity.

Relationship Between ELL Status and Writing Proficiency

Once the structure of the DWA was examined, I answered research questions 2, 3, and 4:

2) What is the relation between ELL status and writing proficiency for ninth-grade students attending public schools in Utah during the years 2004, 2005, 2006, and 2007, and to what extent do student variables, gender, social economic status, and ethnicity independently and cumulatively explain the relationship?

3) To what extent do the school variables, percent of low-income students in a school, percent minority students in a school, size of the school, and mean

ELL status at a school independently and cumulatively explain the relation between ELL status and writing proficiency?

4) To what extent do the district variables, percent low-income students in a district, percent minority students in a district, size of the district, mean ELL status in a district, and whether a district is urban or rural independently and cumulatively explain the relation between ELL status and writing proficiency?

I investigated the relationship between ELL status and writing proficiency of ninth-grade students who attended Utah's public schools during the 2004, 2005, 2006, and 2007 school years. Moreover, I assessed the relative contributions of student, school, and district to the relationship between ELL status and writing proficiency of ninth-grade students. I identified predictor variables at the individual, school, and district level that contribute to the relationship. The predictor variables chosen for this study were variables that have been previously linked to writing proficiency. In addition, I investigated the extent to which the relationship between ELL status and writing proficiency is moderated by ethnicity. These analyses indicated whether there were differences in the relationship between ELL status and writing proficiency by ethnic group. Finally, I examined each of the six traits that are used in the DWA by ethnicity. These analyses indicated whether any of the six traits vary by different cultural groups.

In many investigations of student achievement, researchers have ignored the consequences of aggregated data from students within classrooms and schools (Bryk & Raudenbush, 1992). Because students in the same school will likely be more similar to one another than students from different schools, results often appear to have greater significance than what is justified (Nye, Hedges, & Konstantopoulos, 2000). Many

researchers have statistically analyzed data at the level of the individual, using commonly accepted statistical techniques such as multiple regression or analysis of variance. These techniques require that all observations are independent and error terms should not be correlated. Unfortunately, aggregated data ignores the effect that context may have on subsets of the data. This is certainly the case when working with large data sets involving student achievement data. Violations of independence and correlated error terms occur because there are often classroom, school, and even district effects. The amount of dependency that is observable due to aggregation can be measured using the intraclass correlation. With an increase in the intraclass correlation, the amount of independency decreases, inflating the Type I error rate and resulting in tests of significance that are too liberal (Hedeker, Gibbons, & Flay, 1994). Type I error rates at .05 level can be as high as .17 when group size is 100, and can rise to .70 when the intraclass correlation is .20 (Tabachnick & Fidell, 2007).

Apart from the statistical violations that aggregation of data leads to, aggregating data ignores the general nature of student-achievement data. Classroom effects, school effects, and district effects have long been recognized in educational research. Not all classrooms, schools, and districts are created equal. By wiping out this stark reality and aggregating data, statistical models have been produced that are assumed to apply equally to all contexts (Luke, 2004). For example, what contributes to academic success or failure in one school is assumed to contribute to academic success or failure in another school. Such blanket statements cannot be substantiated.

Because variance in student achievement within schools is often different from the variance between schools, and the variance between schools is often different from

the variance of schools within districts, the analyses of data should take into account this hierarchical nature. Hierarchical linear modeling (HLM) does just this (Bryk & Raudenbush, 1992). HLM has been developed for research designs using data that represent multiple levels and for which there may be different variables available at each level. For example, when working with student, school, and district-level data, ethnicity may contribute at the student level, school size may contribute at the school level, and urban/rural may contribute at the district level. Moreover, there may be cross-level interactions, such as school size interacting with ethnicity, or urban/rural interacting with ethnicity.

HLM is essentially a sophisticated multiple regression analysis in which multiple predictor variables are used to predict an outcome variable. In HLM, regression intercepts (means) and slopes (IV-DV relationships) at one level are allowed to vary between higher level units. This variability is modeled by treating group intercepts and slopes at one level as DVs in the next higher level. With HLM, independence of errors is not required and is often violated. HLM also allows predictor variables at every level of analysis. Higher-level predictors can help in explaining lower-level differences in intercepts and slopes. For example, differences among schools on writing proficiency could be a function of whether districts are urban or rural. Therefore, within-level interactions among predictors, such as school size and school ethnicity, and cross-level interactions, such as school ethnicity and district size, can be modeled. Moreover, unequal sample sizes within each level pose no problems. For these analyses, each year of data was analyzed separately. Finding similar results for each year serves as evidence of generalizability for

the findings, which adds to the body of evidence collected concerning the validity of the DWA.

Measures

Six Traits

Each student's essay is evaluated using the six traits (i.e., voice, organization, ideas, conventions, word choice, and sentence fluency). An essay is assigned a score ranging from 1-to-5 on each trait and a composite score is calculated. Because of the high correlations among the traits, students' composite scores were used as the dependent variable. A description of the six traits and the scoring criteria are in Appendix A.

Student Variables

Student variables are gender, socioeconomic status, ethnicity, and ELL status. Writing proficiency has traditionally been differentiated on the basis of gender, with females outperforming males. Therefore, gender was included and coded as a categorical variable with 0 for male and 1 for female. Writing proficiency also has been differentiated on the basis of social economic status (SES). Therefore, SES was included as a categorical variable and coded on the basis of students' eligibility for a free or reduced lunch program. Students eligible for free or reduced lunch were coded as 1, and all other students were coded as 0. Writing proficiency has been differentiated on the basis of ethnicity. Therefore, ethnicity served as a dichotomous variable with 0 = White and 1 = to a combined Asian, African American, Hispanic, Native American, and

Polynesian. By collapsing across all ethnic groups other than White, the number of students within the two groups were more evenly balanced.

Last, ELL status was a target variable in this study and served as an ordinal variable. Although ELL designations have changed over the past few years, when the data for this study were collected, there was a fairly uniform policy in place. Once a student was enrolled in a school district, the parents of the student completed the Primary Home Langue Other Than English (PHLOTE) form. This information indicated what language was the primary language used in the student's residence. The student was then tested with either the IDEA Proficiency Test (IPT) or the Utah Academic Language Proficiency Assessment (UALPA). Based on the results of either test, the student is given an ELL designation. Students are tested at the end of every year with the same test, and the placement for the following years is determined. The ELL designations and descriptions are:

Pre-Emergent

- Student has limited or no understanding of oral or written English but participates by listening.
- Student may demonstrate comprehension by using a few isolated words or expressions in speech.
- Student typically draws, copies, or responds verbally, or in their native language to simple commands, statements, and questions.
- Student may understand the relationship between oral and written language.

Emergent

- Student begins to understand that written language represents oral language.

- Student understands and responds to basic social conventions, simple questions, simple directions, and appropriate-level text.
- Student participates in classroom routines.
- Student speaks, reads, and writes using single words, short phrases, or simple sentences with support.
- Student has minimal expressive vocabulary.

Intermediate

- Student understands and uses more abstract, academic, and formal language and literacy skills.
- Student participates actively in most social and classroom tasks, using some idioms and more content-specific language in speech and writing.
- Student reads and writes independently for personal and academic purposes, with some persistent errors.
- With some support, student reads and writes about various topics, using different genres for a variety of audiences.

Advanced

- Student has developed proficiency in English language and literacy skills.
- Student may need continued support when engaging in complex academic tasks requiring increasingly academic language.
- Most students at this level attend mainstream or sheltered instruction classes.

Fluent

- Student has exited from the program having achieved fluency and is participating fully in mainstream classes.

- A student is monitored for 2 years after testing as fluent.
- After those 2 years, the student is categorized as an "E" or "exited" from the program.

Because these ELL designations are intended to represent a continuous progression of English language proficiency, I will treat these designations as a continuous variable, with: 1 = Pre-Emergent, 2 = Emergent, 3 = Intermediate, 4 = Advanced, and 5 = Fluent. For all students whose first language is English, I will designate them with a 6. Therefore, use of the 1-to-6 scale is an ordinal variable with increasing English-language proficiency with an increase in number. This variable was centered in the level 1 model.

School Variables

School variables are number of ninth-grade students at a school, the mean ELL status at ninth-grade at a school, and the proportion of ninth-grade students in a school who were receiving free or reduced-price lunch. Although my data are restricted to ninth grade at any given year, I have 4 years of data ranging from 2004 to 2007. Therefore, by analyzing all 4 years separately, I can generalize my results across at least 4 years of schooling.

District Variables

District variables are number of ninth-grade students in a district, the mean ELL status at ninth-grade in a district, the proportion of ninth-grade students in a district who

were low-income, and whether the district is urban or rural. Urban/rural was a dichotomous variable, with 0 = rural and 1 = urban.

Analysis Overview

Level One: Student

The lowest level of the hierarchy models individual student data within school. Student writing proficiency was analyzed as a function of gender (dichotomous variable), English Language Learning (1 to 6 scale, centered), ethnicity (dichotomous variable), and SES (dichotomous variable measuring whether students receive or do not receive free or reduced lunch programs). The specific model used was:

$$Y_{ijk} = \pi_{0jk} + \pi_{1jk}(\text{gender}) + \pi_{2jk}(\text{low-income}) + \pi_{3jk}(\text{White/non-White}) + \pi_{4jk}(\text{ELL-centered}) + \varepsilon_{ijk}$$

Where Y_{ijk} = writing score for individual i in school j in district k;

π_{0jk} = intercept (mean writing score) of school j in district k;

π_{1jk} = linear slope of the relation between gender and writing score in school j in district k;

π_{2jk} = linear slope of the relation between low-income and writing score in school j in district k;

π_{3jk} = linear slope of the relation between White/non-White and writing score in school j in district k;

π_{4jk} = linear slope of the relation between ELL and writing score in school j in district k;

ε_{ijk}= error in writing score for individual i in school j in district k.

Level Two: School

The middle level of the hierarchy models school variation. At this level, I modeled variation across schools for each school-specific regression coefficient according to the size of the school (number of students at ninth grade), the mean ELL status at ninth grade, and the proportion of ninth-grade students in the school who were low-income (receiving free or reduced lunch). The specific model used was:

$$\pi_{0jk} = \beta_{00k} + \beta_{01k}(\#\text{ of students}) + \beta_{02k}(\text{mean ELL}) + \beta_{03k}(\text{proportion low-income}) + \beta_{04k}(\text{proportion non-White}) + r_{0jk}$$

Where π_{0jk} = intercept (mean writing score) of school j in district k;

β_{00k} = mean intercept (mean writing performance) for all schools within district k;

β_{01k} = strength and direction of the relation between number of ninth-grade students at a school and the school's mean writing score for all schools in district k;

β_{02k} = strength and direction of the relation between proportion low-income at ninth grade at a school and the school's mean writing score for all schools in district k;

β_{03k} = strength and direction of the relation between mean ELL status at ninth-grade at a school and the school's mean writing score for all schools in district k;

β_{04k} = strength and direction of the relation between proportion non-White at ninth-grade at a school and the school's mean writing score for all schools in district k;

r_{0jk} = residual for school j in district k.

$$\pi_{1jk\ (gender)} = \beta_{10k} + \beta_{11k}(\#\ of\ students) + \beta_{12k}(mean\ ELL) + \beta_{13k}(proportion\ low\text{-}income) + \beta_{14k}(proportion\ non\text{-}White) + r1_{jk}$$

where π_{1jk} = linear slope of the relation between gender and writing score in school j in district k;

β_{10k} = mean linear slope between gender and writing score for all schools in district k;

β_{11k} = strength and direction of the relation between number of students at ninth grade at a school and relation between gender and writing score for all schools in district k (effect of number of students on relation between gender and writing score for schools in district k);

β_{12k} = strength and direction of the relation between mean ELL status at ninth grade at a school and relation between gender and writing score for all schools

in district k (effect of mean ELL on relation between gender and writing score for schools in district k);

β_{13k} = strength and direction of the relation between proportion low-income at a school and relation between gender and writing score for all schools in district k (effect of proportion low-income on relation between gender and writing score for j schools in district k);

β_{14k} = strength and direction of the relation between proportion non-White at a school and relation between gender and writing score for all schools in district k (effect of proportion low-income on relation between gender and writing score for j schools in district k);

r_{1jk} = residual for school j in district k for relation between gender and writing score.

$$\pi_{2jk\ (low\text{-}income)} = \beta_{20k} + \beta_{21k}(\#\ of\ students) + \beta_{22k}(mean\ ELL) + \beta_{23k}(proportion\ low\text{-}income) + \beta_{24k}(proportion\ non\text{-}White) + r_{2jk}$$

where $\pi_{2jk\ (ELL)}$ = linear slope of the relation between ELL and writing score in school j in district k;

β_{20k} = mean linear slope between ELL and writing score for all schools in district k;

β_{21k} = strength and direction of the relation between number of students at a school and relation between ELL and writing score for all schools in district k

(effect of number of students on relation between ELL and writing score for all schools in district k);

β_{22k} = strength and direction of the relation between mean ELL status at a school and relation between ELL and writing score for all schools in district k (effect of mean ELL status at a school on relation between ELL and writing score for all schools in district k);

β_{23k} = strength and direction of the relation between proportion low-income at a school and relation between ELL and writing score for all schools in district k (effect of proportion low-income on relation between ELL and writing score for all schools in district k);

β_{24k} = strength and direction of the relation between proportion non-White at a school and relation between ELL and writing score for all schools in district k (effect of proportion low-income on relation between ELL and writing score for all schools in district k);

r_{2jk} = residual for school j in district k for relation between ELL and writing score.

$$\pi_{3jk\ (\text{low-income})} = \beta_{30k} + \beta_{31k}(\#\ \text{of students}) + \beta_{32k}(\text{mean ELL}) + \beta_{33k}(\text{proportion low-income}) + \beta_{34k}(\text{proportion non-White}) + r_{3jk}$$

where π_{3jk} = linear slope of the relation between low-income and writing score in school j in district k;

β_{30k} = mean linear slope between low-income and writing score for all schools in district k;

β_{31k} = strength and direction of the relation between number of students at a school and relation between low-income and writing score for all schools in district k (effect of number of students on relation between low-income and writing score for schools in district k);

β_{32k} = strength and direction of the relation between mean ELL status at a school and relation between low-income and writing score for all schools in district k (effect of mean ELL status on relation between low-income and writing score for schools in district k);

β_{33k} = strength and direction of the relation between proportion of low-income students at a school and relation between low-income and writing score for schools in district k (effect of proportion low-income on relation between low-income and writing score for all schools in district k);

β_{34k} = strength and direction of the relation between proportion of non-White students at a school and relation between low-income and writing score for schools in district k (effect of proportion low-income on relation between low-income and writing score for all schools in district k);

r_{3jk} = residual for school j in district k for relation between low-income and writing score.

Level Three: District

The third level of the hierarchy models district variation. At this level, I modeled variation across districts for each district-specific regression coefficient according to the

size of the district (number of students at ninth-grade), the district mean ELL status, the proportion of ninth-grade students in the district who were low-income, and whether the district is urban or rural. To illustrate the specific model that was used, only a representative sample of coefficients is described:

$$\beta_{00k} = \gamma_{000} + \gamma_{001}(\text{\# of schools within district}) + \gamma_{002}(\text{mean ELL}) + \gamma_{003}(\text{proportion low-income within district}) + \gamma_{004}(\text{proportion non-White within district}) + \gamma_{005}(\text{urban/rural}) + U_{00k}$$

where β_{00k} = mean intercept (mean writing score) for all schools within district k;

γ_{000} = mean intercept (mean writing score) across districts;

γ_{001} = mean relation between district size and writing score across districts;

γ_{002} = mean relation between mean ELL status and writing score across districts;

γ_{003} = mean relation between proportion low-income and writing score across districts;

γ_{004} = mean relation between proportion non-White and writing score across districts;

γ_{005} = mean relation between urban/rural status and writing score across districts;

U_{00k} = uniqueness in writing score of district k from all districts.

$$\beta_{01k} = \gamma_{010} + \gamma_{011}(\text{district size}) + \gamma_{012}(\text{proportion ELL within district}) + \gamma_{013}(\text{proportion low-income within district}) + \gamma_{014}(\text{proportion non-White within district}) + \gamma_{015}(\text{urban/rural}) + U_{01k}$$

where β_{01k} = strength and direction of the relation between number of students at a school and the school's mean writing score for all schools in district k;

γ_{010} = mean relation between number of students within a district and writing score across districts;

γ_{011} = the effect of district size on the relation between number of students in a district and writing score across districts;

γ_{012} = effect of proportion ELL within district on relation between number of students in a district and writing score across districts;

γ_{013} = effect of proportion low-income within district on relation between number of students in a district and writing score across districts;

γ_{014} = effect of proportion non-White within district on relation between number of students in a district and writing score across districts;

γ_{015} = effect of urban/rural status of a district on relation between number of students in a district and writing score across districts;

U_{01k} = uniqueness in writing score of district k from all districts due to urban/rural

Relationship Between ELL Status and Ethnicity and the Six Traits

I used a multivariate analysis of variance (MANOVA) to answer research my last research question:

5) To what extent do each of the six traits vary with ELL status and ethnicity, and are there interactions between ELL status and ethnicity on any of the six traits?

In this analysis, I used the two categorical variables of ELL status and ethnicity as independent variables, and the six trait scores were used as the dependent variables. Thus, the analysis was an ELL status (5 levels) x Ethnicity (6 levels) MANOVA with six dependent variables. I tested for main effects for ELL status and ethnicity (i.e., does scoring on the six traits vary by ELL status or ethnicity), and for interactions between ELL status and ethnicity (i.e., does change in scores on the six traits over the levels of ELL status depend on ethnicity). For example, Hispanic students with an emergent ELL status may score higher on voice than African-American students with an emergent ELL status.

CHAPTER 4

RESULTS

Research Question One

1) Does the DWA provide reliable scores and valid interpretations of those scores concerning writing proficiency for students in general and for specific groups of students based on ELL status and ethnicity?

To answer this question, I followed five guidelines as described by Osterlind (2006). First, I reviewed the published information about the DWA that was provided by the test's author. Second, for each year that I have DWA data, I calculated summary statistics for each of the six traits. These statistics provided essential information concerning the distributions of the scores for each of the six traits for all students. Third, summary statistics from the DWA were compared to similar summary statistics from The National Assessment for Education Progress (NAEP) for Writing 1998, 2002, and 2007. I looked for similar trends in the data across the two tests that include gender, ethnicity, and social economic status. Fourth, I examined the reliability of the scores from the DWA. I examined the correlations among the six traits, calculated Cronbach's α for all

students and for groups at the various levels of ELL. Fifth, to evaluate validity of the DWA, I took a four-pronged approach. In this approach, I collected various kinds of empirical evidence and theoretical rationales to evaluate the proposed interpretations of the test scores in light of the purpose of the DWA.

The first prong of my examination of validity concerned the underlying construct that is measured by the DWA. That construct is writing skill, which is defined by the six traits: voice, organization, ideas, conventions, word choice, and sentence fluency. To provide support for the construct validity of the six traits, I described what is the intended meaning of each and examined a variety of writing assessments for evidence of these same traits.

Second, I examined the extent to which the six traits are used in instruction and assessment in schools across the United States. The ultimate goal was to demonstrate that the six traits are widely recognized as defining characteristics of writing proficiency and, hence, are widely used in instruction and assessments.

The third prong of my examination of validity was to analyze the internal structure of the DWA. Using 1 year of data, I first conducted an exploratory factor analysis to investigate the dimensionality of the test. In many writing assessments, there are often two dimensions identified, one dealing with the mechanics or conventions of writing and one dealing with the content and how that content is expressed. The purpose of the factor analysis was to uncover whether this dimensionality exists for the DWA. This analysis was followed by a confirmatory factor analysis using structural equation modeling (SEM). Along with the structural equation modeling, I examined whether the

same identified dimensionality of the DWA is present for the targeted subgroups of interest in this study.

Finally, the fourth prong of my examination of validity was to analyze data from other years, using the same methods just described. The degree to which evidence of validity based on 1 year of data generalized to other years is reflective of the test's overall validity.

Published Information About the DWA

From 2004 to 2007, Pearson Publication Company was contracted by the Utah State Office of Education to score the DWA; however, Pearson Publication did not have input as to the content of the DWA. That is, use of the six traits in the DWA was a decision made by the Utah State Office of Education, and Pearson Publication was provided with the task of scoring student performance. Therefore, the published information from Pearson Publication is focused almost entirely on the scoring procedure and little information is provided on validity or other psychometric qualities of the DWA. The following scoring information has been extracted from the technical report produced by the Pearson Publication Company (2008). Athough the report is for the 2008 data, the same or similar procedures for scoring essays were followed for the 2004 to 2007 data.

Scorers were distributed across the country and scored their assigned essays using an electronic network that allowed them to view and score scanned images of essays. Scorers, some of whom were Utah educators, were asked to commit to 15 to 20 hours per week. All essays received an initial read, and 20% of them were scored a second time to monitor the quality of the scores. To qualify to score the DWA, scorers were required to

complete an online training course that consisted of several modules. Each module had to be completed before moving on to the next. The modules were: project overview, definition and explanation of USOE's DWA traits, rubric/scoring guide, scoring versus grading, reader bias, anchor papers, condition codes and alert papers, and practice papers. After the modules were completed, scorers were required to qualify to score the DWA. Scorers were given four sets of 10 qualification papers and were required to achieve at least 60% exact agreement on two of the four sets.

Scoring directors at Pearson used scorer performance reports, including interrater reliability, frequency distribution, validity, calibration, and backreading, to monitor accuracy and scoring trends. The interrater reliability report indicates how many times scorers are in exact agreement or assign adjacent or nonadjacent scores. USOE provided approved essays that were used to maintain consistency in scoring over time. These essays were prescored and then interspersed with other student essays, and were indistinguishable from other student essays. The scoring on these essays was then checked throughout the scoring process, and decisions were made as to how well scorers were remaining consistent in their scoring.

To provide feedback on the accuracy and consistency of scoring, expert scorers "backread" approximately 5% of the essays to which scorers have assigned scores. These expert scorers then provided feedback to the scorers on their accuracy and consistency of their scores. If a scorer fell below certain standards, he or she was not allowed to continue scoring.

The interrater reliability for the 2008 DWA is reproduced here. The original information appeared in the Pearson Publication Company Technical Report (2008).

Interrater consistency is reported as the percentage of exact and adjacent score point agreement. Adjacent agreement means that scorer's trait score was within one point, exact agreement means that scorers gave the same trait score.

According to the Pearson Technical Report for 2008, the combined rate of exact and adjacent score-point agreements was greater than 98% for each of the six traits. Pearson states that, "This level of inter-rater scoring agreement is consistent with PSC professional scoring standards and provides a more appropriate index of interscorer consistency than the Pearson correlation" (p. 33).

Summary Statistics, Correlations, and Reliability

For each year that I have DWA data, I calculated summary statistics for each of the six traits. These statistics provided essential information concerning the distributions of the scores for each of the six traits for all students, the correlations among the traits, and the reliability of the scores (see Tables 1 through 3).

Tables 4 through 7 show the descriptive statistics for 2004, 2005, 2006, and 2007, respectively. The means and standard deviations for each of the traits for each of the years are stable, ranging from 3.45 to 3.69 for the means, and from .67 to .77 for the standard deviations. This stability is reflected in the high Cronbach Alphas for each year: .94 for 2004 and 2005, .96 for 2006 and 2007. Although there are statistical differences between several of the traits for each year, the effect size is small (eta squared < .001). Also, the Pearson correlations among the traits are high, ranging between .50 and .92, with intraclass correlations ranging between .72 and .80. Finally, measures of skewness

Table 1

Interrater Reliability Hit Rates for Reader 1 and Reader 2

Trait	Agreement	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Ideas	> 1 point	48	.66	48	.66
	Adjacent	2503	34.33	2551	34.99
	Exact	4740	65.01	7291	100.00
Organization	> 1 point	48	.66	48	.66
	Adjacent	2442	33.49	2490	34.15
	Exact	4801	65.85	7291	100.00
Voice	> 1 point	50	.69	50	.69
	Adjacent	2559	35.10	2609	35.78
	Exact	4682	64.22	7291	100.00
Word Choice	> 1 point	51	.70	51	.70
	Adjacent	2645	36.28	2696	36.98
	Exact	4595	63.02	7291	100.00
Sentence Fluency	> 1 point	85	1.17	85	1.17
	Adjacent	2733	37.48	2818	38.65
	Exact	4473	61.35	7291	100.00
Conventions	> 1 point	98	1.34	98	1.34
	Adjacent	2781	38.14	2879	39.49
	Exact	4412	60.51	7291	100.00

Table 2

Interrater Reliability Hit Rates for Reader 1 and Read Behind

Trait	Agreement	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Ideas	> 1 point	1	.34	1	.34
	Adjacent	72	24.24	73	24.58
	Exact	224	75.42	297	100.00
Organization	> 1 point	0	0	0	0
	Adjacent	73	24.58	73	24.58
	Exact	224	75.42	297	100.00
Voice	> 1 point	0	0	0	0
	Adjacent	89	29.97	89	29.97
	Exact	208	70.03	297	100.00
Word Choice	> 1 point	0	0	0	0
	Adjacent	78	26.26	78	26.26
	Exact	219	73.74	297	100.00
Sentence Fluency	> 1 point	0	0	0	0
	Adjacent	95	31.99	95	31.99
	Exact	202	68.01	297	100.00
Conventions	> 1 point	4	1.35	4	1.35
	Adjacent	98	33.00	102	34.34
	Exact	195	65.66	297	100.00

Table 3

Interrater Reliability Hit Rates for Reader 2 and Read Behind

Trait	Agreement	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Ideas	> 1 point	1	.50	1	.50
	Adjacent	67	33.33	68	33.83
	Exact	133	66.17	201	100.00
Organization	> 1 point	0	0	0	0
	Adjacent	60	29.85	60	29.85
	Exact	141	70.15	201	100.00
Voice	> 1 point	1	.50	1	.50
	Adjacent	66	32.84	67	33.33
	Exact	134	66.67	201	100.00
Word Choice	> 1 point	1	.50	1	.50
	Adjacent	81	40.30	82	40.80
	Exact	119	59.20	201	100.00
Sentence Fluency	> 1 point	1	.50	1	.50
	Adjacent	78	38.81	79	39.30
	Exact	122	60.70	201	100.00
Conventions	> 1 point	3	1.49	3	1.49
	Adjacent	72	35.82	75	37.31
	Exact	126	62.69	201	100.00

Table 4

Descriptive Statistics for 2004 Data (N = 34,108)

	Ideas	Organization	Voice	Word Choice	Sentence Fluency	Conventions
Ideas		.87	.67	.71	.74	.70
Organization			.67	.72	.76	.72
Voice				.69	.65	.59
Word Choice					.77	.71
Sentence Fluency						.81
Mean (SD)	3.58 (.73)	3.58 (.74)	3.56 (.67)	3.49 (.68)	3.57 (.73)	3.53 (.76)
Skewness (SE)	-.43 (.01)	-.38 (.01)	.06 (.01)	.12 (.01)	-.24 (.01)	-.25 (.01)
Kurtosis (SE)	.28 (.03)	.20 (.03)	.00 (.03)	.09 (.03)	.04 (.03)	.00 (.03)

All correlations are significant at the .01 level (two-tailed).

Cronbach Alpha = .94

Intraclass correlation = .72

Table 5

Descriptive Statistics for 2005 Data (N = 26,811)

	Ideas	Organization	Voice	Word Choice	Sentence Fluency	Conventions
Ideas		.86	.68	.72	.73	.71
Organization			.69	.72	.77	.74
Voice				.71	.69	.65
Word Choice					.78	.74
Sentence Fluency						.83
Mean	3.58	3.55	3.69	3.45	3.53	3.48
(SD)	(.72)	(.74)	(.70)	(.68)	(.74)	(.77)
Skewness	-.45	-.38	-.16	-.01	-.22	-.34
(SE)	(.02)	(.02)	(.02)	(.02)	(.02)	(.02)
Kurtosis	.31	.21	.07	.24	.06	.11
(SE)	(.03)	(.03)	(.03)	(.03)	(.03)	(.03)

All correlations are significant at the .01 level (two-tailed).

Cronbach Alpha = .94

Intraclass correlation = .73

Table 6

Descriptive Statistics for 2006 Data (N = 33,627)

	Ideas	Organization	Voice	Word Choice	Sentence Fluency	Conventions
Ideas		.92	.77	.77	.79	.77
Organization			.77	.78	.81	.80
Voice				.82	.76	.74
Word Choice					.84	.82
Sentence Fluency						.90
Mean (SD)	3.52 (.72)	3.52 (.72)	3.66 (.68)	3.54 (.68)	3.54 (.72)	3.51 (.74)
Skewness (SE)	-.38 (.01)	-.41 (.01)	-.22 (.01)	-.11 (.01)	-.30 (.01)	-.34 (.01)
Kurtosis (SE)	.26 (.03)	.37 (.03)	.19 (.03)	.25 (.03)	.32 (.03)	.32 (.03)

All correlations are significant at the .01 level (two-tailed).

Cronbach Alpha = .96

Intraclass correlation = .80

Table 7

Descriptive Statistics for 2007 Data (N = 36,381)

	Ideas	Organization	Voice	Word Choice	Sentence Fluency	Conventions
Ideas		.91	.79	.80	.79	.78
Organization			.78	.79	.79	.79
Voice				.80	.75	.74
Word Choice					.86	.84
Sentence Fluency						.88
Mean	3.55	3.57	3.68	3.54	3.56	3.52
SD	.75	.76	.71	.72	.73	.75
Skewness (SE)	-.30 (.01)	-.40 (.01)	-.32 (.01)	-.23 (.01)	-.35 (.01)	-.35 (.01)
Kurtosis (SE)	.09 (.03)	.20 (.03)	.24 (.03)	.22 (.03)	.23 (.03)	.23 (.03)

All correlations are significant at the .01 level (two-tailed).

Cronbach Alpha = .96

Intraclass correlation = .80

and kurtosis indicate that with only two exceptions, the distributions for each of the traits for each of the years is slightly negatively skewed and clustered more than a normal distribution. However, with the large sample sizes involved, the slight deviations from normality in the distributions do not make a substantive difference in the analyses planned for this study (Tabachnick & Fidell, 2007). The high correlations among the traits' scores justifies the use of the total score as the dependent variable in all analyses.

To measure the degree to which each trait is correlated with other traits, I calculated the corrected item-total correlation and the squared multiple correlation. The corrected item-total correlation compares each individual item to a scale composed of the other items, and the squared multiple correlation indicates the amount of variability in the scale of items that is accounted for by the item that is not included in the scale. As can be seen from Table 8, for each year of the DWA, the corrected item-total correlations indicate that each individual item is strongly correlated with the other items, and each item can account for a majority of the variability in the scale of the other items. Therefore, a high score on one item was strongly associated with high scores on the other items, and conversely, a low score on one item was strongly associated with low scores on the other items. Organization has the highest correlation with the scale of other items consistently across the 4 years of data. This could have implications for future scoring of the DWA in that if only organization were scored, about 83% of the variability in the other scores can be accounted for.

Table 8

Corrected Item-Total Correlation and Squared Multiple Correlation for Each Trait for Each Year

Year	Trait	Corrected Item-Total Correlation	Squared Multiple Correlation
2004	Ideas	.85	.79
	Organization	.86	.80
	Voice	.73	.55
	Word Choice	.82	.68
	Sentence Fluency	.86	.76
	Conventions	.80	.69
2005	Ideas	.83	.76
	Organization	.86	.79
	Voice	.76	.59
	Word Choice	.83	.69
	Sentence Fluency	.86	.77
	Conventions	.83	.72
2006	Ideas	.88	.86
	Organization	.89	.87
	Voice	.83	.72
	Word Choice	.88	.79
	Sentence Fluency	.90	.86
	Conventions	.88	.84
2007	Ideas	.88	.84
	Organization	.89	.85
	Voice	.83	.71
	Word Choice	.89	.81
	Sentence Fluency	.89	.83
	Conventions	.88	.81

Comparison With NAEP

The National Assessment of Educational Progress for Writing (aka, The Nation's Report Card) is a report produced by the National Center for Education Statistics of the U.S. Department of Education. The report provides the results of national assessments of students' achievement in writing. Other reports focus on achievement in other academic areas, such as reading, mathematics, and science. The reports are produced approximately every 4 years, and for the purposes of the present study, reports from 1998, 2002, and 2007 are used. Each report is based on large, nationally representative samples. For example, the 2007 report included a sample of more than 165,000 eighth and 12^{th}graders. Earlier reports have included fourth-grade students; however, this group was not represented in the 2007 report. Participating students write three essays (i.e., to inform, to persuade, and to tell a story) for a variety of audiences. The essays are then evaluated using primary trait assessment in which a single holistic score is assigned to each essay from each genre. Results are reported using a 0-300 scale and are reported as percentages of students performing at above basic, proficient, and advanced levels. Essays are scored using a rigorous procedure that meets high standards of validity and reliability. For a complete description of the scoring procedure, please see Appendix A of the 2002 Report (NAEP, 2002).

The NAEP reports for writing provide a breakdown of the results for a variety of categories, for example, gender, ethnicity, and income. The purpose of the following section is to compare results from the NAEP reports for 1998, 2002, and 2007 to the results from the DWA for years 2004 to 2007 to identify similar trends in the two tests. If results from the smaller scale DWA conform with results from the large-scale NAEP, this

would indicate that the kinds of empirical evidence provided by both tests would support similar interpretations of the scores. Thus, the two sources of evidence point to similar valid inferences about writing skill (Messick, 1989).

Gains in Writing Performance

The results from all 3 years of the NAEP indicate that the 2007 results were better than the previous years for both grade levels. For the eighth grade, the average writing score was 3 points higher in 2002 and 6 points higher than in 1998. Twelfth-grade students performed 5 points higher than in 2002 and 3 points higher than in 1998.

For the DWA, writing performance for all students across the 4 years was better in 2007 than in any of the 3 preceding years, $F(3, 130923) = 7.105$, $p < .001$, eta squared < .001; however, the effect size was very small (see Table 9). There were no differences for years 2004 to 2006. Therefore, although there was an increase in writing skill between years 2006 and 2007 for ninth-grade students in Utah, the gains in writing have been minimal.

Table 9

Overall Gains in Writing Performance for DWA

	Score
2004	21.32 (3.77)
2005	21.29 (3.84)
2006	21.29 (3.90)
2007	21.40 (4.05)

Writing Performance by Gender

The results from all 3 years of the NAEP data (1998, 2002, and 2007), for both girls and boys in the 8th and 12th grades indicate that girls continue to outperform boys in writing. While both groups have increased their performance in writing over the different testing periods, the gap between the genders remains consistent. In the 8th grade, girls continued to outperform boys by about 20 points. In the 12th grade, the most recent testing pointed to an 18-point gap between girls and boys.

The DWA test results from 2004-2007 show that girls outperformed boys, $F(1, 130910) = 721.146$, $p < .001$, eta squared = .005, although the effect size was small. This was true for each for each ethnic group as well (see Table 10). Both the DWA and NAEP results indicate that the gap between girls and boys continues to exist.

Table 10

Writing Performance by Gender and Ethnicity for 2004-2007

	Female	Male
Asian	22.84 (3.73)	21.43 (4.02)
African-American	20.76 (3.80)	18.90 (4.09)
White	22.37 (3.49)	20.90 (3.87)
Hispanic	20.14 (3.93)	18.67 (4.22)
Indian	19.79 (3.91)	17.90 (4.28)
Pacific Islander	21.61 (3.57)	20.09 (3.86)

Gains in Writing Performance by Ethnicity

The NAEP results show that White, Black, and Asian/Pacific Islander eighth-grade students had higher average writing scores in 2007 than in 2002 and 1998. The average score for Hispanic students was higher in 2007 than in both previous assessments. Asian and Pacific Islander students scored higher than in 2002, but in comparison to 1998, it was not statistically significant. There were no significant changes in the average writing scores for American Indian/Alaska Native students compared to previous years. Asian/ Pacific Islanders continue to outperform all other ethnic groups. Twelfth-grade White, Black, and Asian/Pacific Islander students scored higher in 2007 than in both previous assessment years. Hispanic students did not see any significant change. Black and Asian/Pacific Islander students scored higher than in 2002, but apparent changes in comparison to 1998 were not statistically significant. Scores for Hispanic and American Indian/Alaska Native students showed no significant change in comparison to previous assessments.

The DWA results show that Asian and White students performed about the same while outperforming all other ethnic groups across the four years (see Table 11). One exception, however, was in 2005 when Pacific Islanders were on par with both Asian and White students. Pacific Islander students tended to perform on par with African American students and to outperform Hispanic and American Indian students. The results for African American students were mixed. For 2004 and 2007, they outperformed Hispanic and Indian students, but for 2005 and 2006, they were on par with Hispanic and American Indian students. Hispanic and American Indian students tended to have the lowest performance of all ethnic groups across the 4 years. Across the 4 years, there were

Table 11

Writing Performance by Ethnicity

	2004	2005	2006	2007
Asian	21.91 (4.00)	21.94 (3.93)	22.11 (3.79)	22.43 (4.03)
White	21.64 (3.64)	21.52 (3.72)	21.59 (3.78)	21.71 (3.90)
Pacific Islander	20.50 (3.74)	21.32. (3.91)	20.94 (3.75)	20.77 (3.76)
African-American	19.73 (3.79)	19.31 (4.26)	19.53 (4.04)	20.26 (4.17)
Hispanic	19.09 (3.81)	19.51 (4.19)	19.28 (4.08)	19.60 (4.41)
Indian	18.44 (4.02)	18.79 (4.01)	18.81 (4.14)	19.30 (4.54)

small fluctuations in scores for the different ethnic groups; however, effect sizes were very small (eta squared < .008).

Both the NAEP and the DWA show improvements in the scores of racial and ethnic groups, although the gains in scores for ninth-grade students in Utah were small. With the exception of Asian students, significant gaps continue to exist between the writing scores of White students and other racial/ethnic groups. There were no significant changes in score gaps between White and Black students or White and Hispanic students compared to previous assessment years. The three groups that tend to have the lowest performance were the Hispanic, American Indian, and African American students.

Writing Performance by Income

The NAEP reports for 1998, 2002, and 2007 show that eighth-grade students eligible for free lunch scored lower than those eligible for reduced-price lunch. Both groups scored lower on average than the students who were not eligible. A 25-point gap exists between students who were eligible for free lunch and those who were not. Twelfth-grade students eligible for free lunch scored lower than those eligible for reduced-price lunch. Both groups scored lower on average than the students who were not eligible. A 19-point gap exists between students who were eligible for free lunch and those who were not.

For each year and across all 4 years of the DWA, students not receiving free or reduced lunch outperformed students who received free or reduced lunch, and the effect size was moderate, $F(1, 130899) = 6134.97$, $p < .001$, eta squared = .045 (see Table 12). The results from both assessments are very clear about the relation between writing

Table 12

Writing Performance by Income

	Free/Reduced Lunch	No Free/Reduced Lunch
2004	19.84 (3.88)	21.93 (3.55)
2005	20.03 (3.83)	21.80 (3.64)
2006	19.98 (3.91)	21.71 (3.77)
2007	20.10 (4.26)	21.90 (3.85)

performance (as well as most other areas of academic areas) and income. Students who are disadvantaged economically perform considerably lower in comparison to those who are not eligible for government sponsored meals.

Writing Performance by ELL Status

Although the NAEP does not provide results by ELL status, this information is available for the 4 years of data from the DWA. These data are directly relevant to the purpose of the present study. Therefore, the results are given here. Table 13 shows that except for 2004, during which pre-emergent students scored much higher than for the other years, the pattern of writing performance for the ELL statuses was expected. The less proficient students are with English language as measured by the UALPA, the less proficient they are with writing skill. Moreover, ELL students who are designated as exited are closely matched in their writing proficiency with non-ELL students.

Table 13

Writing Performance by ELL Status

	2004	2005	2006	2007
Pre-emergent	17.64 (4.08)	13.69 (5.10)	14.74 (4.98)	13.93 (5.23)
Beginning	16.53 (3.60)	16.00 (3.74)	16.61 (3.87)	16.75 (4.14)
Intermediate	17.86 (3.17)	18.00 (3.76)	18.20 (3.38)	18.64 (4.05)
Advanced	19.66 (3.35)	20.23 (3.51)	20.28 (3.75)	20.40 (3.76)
Exited	21.35 (3.58)	21.76 (3.58)	21.03 (3.57)	21.61 (3.86)
Non-ELL	21.54 (3.68)	21.44 (3.75)	21.50 (3.81)	21.60 (3.95)

For descriptive statistics for each year of the DWA data, with a break down ethnic group, male and female, and ELL status, see Appendices B through F.

Validity

I used a four-pronged approach to examine the validity of the DWA. With the first prong, I examined the underlying construct that is measured by the DWA, writing skill, which is defined by the six traits: voice, organization, ideas, conventions, word choice, and sentence fluency. I examined the meaning of each trait and compared these meanings with characteristics of writing used by other writing assessments. For the second prong, I examined the extent to which the six traits are used in instruction and assessment in schools across the United States. I next examined the internal structure of the DWA, using both exploratory factor analysis and confirmatory factor analysis. Finally, I analyzed the DWA data for each year separately, using the methods described and compared results across the years to identify evidence of validity based on how well results generalize across the 4 years.

Comparison of Six Traits With Other Writing Constructs

Several methods have been developed to assess writing. Discrete-point assessment measures students' ability to recognize discrete writing skills, such as syntax and orthography, through the use of multiple choice tests or correction of a writing sample. Frequency-count assessment involves the analysis of writing on the basis of subject-verb agreement, comma faults, number of syllables in a word, words in a sentence, sentences in a paragraph, and total words. Scores such as the Flesch Reading

Ease or the Flesch-Kincaid Grade Level Score are examples of this type of assessment. Holistic assessment provides a single numerical or letter score that is assigned to a writing sample and reflects overall writing quality. The writing assessment used by Educational Testing Services (ETS) is an example of this type of assessment. Primary-trait assessment is similar to holistic assessment, except that specific kinds of writing are assessed, such as descriptive, persuasive, or informational. The writing assessments used by NAEP, which include persuasive, narrative, and informational writing prompts, are examples of primary trait assessment. Finally, analytic assessment identifies specific characteristics of writing, and each characteristic is scored. The DWA represents this type of assessment.

Of these five types of writing assessment, only the last three assess actual writing skill. Therefore, I compared the constructs underlying the six traits with the constructs represented by another analytic assessment, the Six-Subgroup Quality Scale (SSQS), and by a holistic assessment, the College Board SAT Writing Exam. The former of these was developed by Ransdell and Levy (1996) and has been used extensively in their research, as well as by Hacker (2010). The latter is used widely by ETS for college admissions throughout the United States. The purpose of this comparison was to examine whether the writing constructs as measured by the DWA are similar to the writing constructs as measured by two other established writing assessments. In other words, is writing skill as conceptualized by the DWA similarly conceptualized by other writing assessments?

The SSQS is an analytic assessment that uses 13 writing constructs, each rated on a 1-to-5 scale. The 13 constructs are grouped into six categories: word choice and arrangement, technical quality, engagement in content, purpose/audience/tone,

organization and development, and style. On college samples, the SSQS has proven to be reliable, with interrater r's ranging from .80 to .90 (Ransdell & Levy, 1996). Hacker (2010), using a college freshman/sophomore sample found an interrater r of .77. The SSQS predicts Nelson-Denny reading comprehension scores and reliably discriminates between basic and advanced writers (Ransdell & Levy, 1996).

The College Board SAT Writing Exam is assessed using a holistic approach in which a single score of 1-to-6 is assigned to an essay, with a score representing the degree of mastery of five criteria (College Board, 2006). These criteria are point of view and critical thinking, organized and clearly focused, use of language and vocabulary, sentence structure, and mechanics. Interrater reliability estimates have ranged from .77 to .81.

Table 14 shows the constructs, their descriptions, and the matching of constructs across the three methods of writing assessment. The three writing assessments are closely matched in three areas: word choice/words/vocabulary, conventions/mechanics/ mechanics, and organization/organization and development/organization. The three assessments are also matched on ideas/content/ideas; however, the SAT also includes an element of organization and coherence of ideas and therefore overlaps with the organization criterion from the DWA and SSQS. The three assessments are matched on sentence fluency/style/sentence structure, although the SSQS also contains a focus on the author being daring in the use of unique ideas, which overlaps to some degree with voice in the DWA and point of view in the SAT. The one area in which there is some divergence among the three assessments is voice/purpose.audience.tone/point of view. Voice in the DWA focuses on whether there is an identifiable person behind the writing.

Table 14

Comparison of Writing Constructs Used in the DWA, SSQS, and SAT Assessments

DWA		SSQS		SAT	
Construct	Description	Construct	Description	Construct	Description
Word Choice	Language is used that is rich, colored, precise, and moves and enlightens the reader.	Words: Choice and arrangement	Sentence meanings are clearly expressed	Vocabulary	The writing exhibits skillful use of language, using a varied, accurate, and apt vocabulary.
Conventions	Mechanical correctness of the piece; spelling, grammar and usage, paragraphing, use of capitals and punctuation.	Mechanics	Correct tenses, grammar, and spelling are used.	Mechanics	The writing is free of most errors in grammar, usage, and mechanics.
Voice	A unique perspective of the writer comes through with honesty, conviction, integrity, and believability.	Purpose/ Audience/ Tone	A definite statement of purpose is stated, and the language and tone used are appropriate for a specific audience.	Point of View	The writing effectively and insightfully develops a point of view on the issue and demonstrates outstanding critical thinking, using clearly appropriate examples, reasons, and other evidence to support its position.

Table 14 Continued

DWA		SSQS		SAT	
Construct	Description	Construct	Description	Construct	Description
Sentence Fluency	Rhythm and flow of the language, the sound of word patterns, and the way in which the writing plays to the ear, not just to the eye.	Style	Sentence structures are concise and to the point with no run-on sentences, and the author has been daring in the use of unique ideas	Sentence Structure	The writing demonstrates meaningful variety in sentence structure.
Ideas	Heart of the message, the content of the piece, the main theme, with details that enrich and develop that theme.	Content	The ideas expressed engage the reader and present alternative points of view.	Ideas	The writing is well-organized and clearly focused, demonstrating clear coherence and smooth progression of ideas.
Organization	Internal structure, the thread of central meaning, the logical and sometimes intriguing pattern of the ideas.	Organization and Development	Support and elaboration have been given, a sense of completeness is provided, and proper paragraphing has been used.	Organization	The writing is well-organized and clearly focused, demonstrating clear coherence and smooth progression of ideas.

That is, the unique qualities of the author stand out in the writing. Purpose.audience.tone of the SSQS includes some elements of voice but focuses more on whether there is a clear purpose stated in the writing. Finally, point of view from the SAT includes the unique qualities as expressed by the author, but it also includes whether the author has shown critical thinking with the use of examples, reasons, and evidence.

Although arriving at the definitive answer to the question of whether writing skill as conceptualized by the DWA is similarly conceptualized by the SSQS and SAT would involve scoring of a large corpus of essays using all three assessments (a procedure that is not possible for the present study), an analysis of the criteria used to conceptualize and measure writing skill across the three methods of assessment can lead some support to the construct validity of the DWA. The ways in which writing is conceptualized and measured by the DWA have clear connections to the SSQS and the SAT. Three of the six characteristics of writing contained in the DWA match closely with characteristics of writing identified in the other two writing assessments. Two of the six match across the three assessments; however, the SSQS and SAT contain additional elements, but those elements overlap with other characteristics contained in the DWA. Only with voice in the DWA is there some divergence from the SSQS and SAT. There are some areas of overlap with purpose.audience.tone from the SSQS and point of view from the SAT, but the major focus of voice appears to be whether there is an identifiable writer behind the writing. This characteristic of writing may not be desirable across all types of genre (e.g., informational text), and therefore, this may be the reason why it is omitted from the SSQS and SAT. In sum, the definition of the construct of writing has more in common across the three writing assessments than what differs.

Acknowledged Use of Six Traits

Throughout our country, the assessment of writing is starting to align more and more with statewide testing or the state standards. The direct assessment of writing is included in many large-scale testing programs, including 31 state assessment programs (Goertz & Duffy, 2001). In the case of the six-trait model, every state is using it in at least one or more of their school districts (Spandel, 2005). Because it has been found to be effective in teaching students not only how to use the criteria to consistently evaluate their own writing (James, Abbott, & Greenwood, 2001), but also to improve their own perception of their writing skills (Isernhagen & Kozisek, 2000), the model is being used by many teachers from South America to Africa and the Far East (Spandel, 2005).

Internal Structure of the DWA

As a starting point in examining the internal structure of the DWA, I used Hacker's (2010) exploratory factor analysis of the SSQS (Ransdell & Levy, 1996). The principal axis factoring of the six subgroups, using varimax rotation with Kaiser normalization showed that the six subgroups loaded on two factors and accounted for 53% of the variability. Content, style, and organization loaded on the first factor, and word choice and mechanics loaded on the second factor. Purpose did not have a sufficiently large loading to be included in the analysis. Relying on the previously discussed analysis of the three writing assessments, the first factor of the SSQS aligns with voice, ideas, and organization of the DWA, and the second factor aligns with

sentence fluency, conventions, and word choice of the DWA. Therefore, I conducted an exploratory factor analysis of the 2004 DWA data with a two-factor structure in mind.

Principal axis extraction with varimax rotation and Kaiser normalization was used with two factors. The Kaiser-Meyer-Olkin measure of sampling was acceptable at .89. The two factors accounted for 76.68% of the variability. The rotated factor matrix appears in Table 15. Sentence fluency, conventions, and word choice loaded strongly on Factor 1 and ideas and organization loaded strongly on Factor 2. Voice, however, loaded equally well on both factors. With the exception of voice, the two factors of the DWA strongly resemble the two factors identified in the exploratory factor analysis of the SSQS, providing some support that the two assessment methods are measuring related writing constructs.

Table 15

Rotated Factor Matrix of Six Traits for 2004 Data With Two Factors

	Factors	
	Factor 1	Factor 2
Sentence Fluency	.81	.45
Conventions	.73	.44
Word Choice	.68	.50
Voice	.53	.52
Ideas	.46	.82
Organization	.50	.79

Although the number of observed variables is small, I conducted an exploratory factor analysis using three factors. The three-factor model was pursued to identify the loading of voice more clearly. With principal axis factoring, varimax rotation, and Kaiser normalization, the three-factor model accounted for 79.83% of the variability, with 28.98% of variance accounted for by Factor 1, 28.68% of the variance accounted for by Factor 2, and 22.17% of the variance accounted for by Factor 3. The Kaiser-Meyer-Olkin measure was .89. The rotated factor matrix appears in Table 16.

Organization and ideas load on Factor 1, sentence fluency and conventions load on Factor 2, and word choice and voice load on Factor 3. The validity of this three-factor structure was supported by conducting the same exploratory factor analysis with each of the other three years of DWA data. With slight variations in factor loadings, the three-

Table 16

Rotated Factor Matrix of Six Traits for 2004 Data With Three Factors

	Factors		
	Factor 1	Factor 2	Factor 3
Organization	.75	.42	.37
Ideas	.75	.39	.39
Sentence Fluency	.40	.72	.41
Conventions	.39	.72	.33
Word Choice	.36	.50	.63
Voice	.41	.32	.60

factor solution was supported in each. This three-factor structure provides a more parsimonious rubric on which to measure writing quality for persuasive essays. Moreover, because of the high correlations that were shown among the six traits for each year of data, a more parsimonious rubric may be warranted.

Ideas and organization both load on one factor. Examining the definitions for these two constructs and the bases on which they are assessed, an appropriate label for this factor would be *patterned ideas*. The ideas of an essay are of course the heart of the essay; however, how the ideas are organized, structured, or patterned within the essay can have critical impact on the meaning. Therefore, the ideas contained within an essay and how those ideas are patterned or structured throughout the essay would define this factor.

Conventions and sentence fluency both load on a second factor. The primary characteristics of sentence fluency are described as the flow of language and sound of word patterns. Appropriate placement of commas, proper use of grammar and usage, paragraphing, and punctuation (i.e., conventions) can strongly influence that flow and sound of language. An appropriate label for this factor could be *fluency components*, that is, the flow of language within an essay and the conventions that contribute to that flow.

Last, word choice and voice both load on the third factor. Word choice is defined as the use of language that is rich, colored, and moves the reader, and voice is defined as the unique perspective of the writer being evident in the writing. The unique writer's perspective is expressed through his or her choice of language used in the essay. An appropriate definition for this factor would be the choice of words in an essay that reveals the writer's unique perspective on the focal topic of the essay. A term to describe this factor would be *writer's perspective*.

To lend further construct validation to the three-factor model produced by the exploratory factor analyses for each year of data, I conducted confirmatory factor analyses using structural equation modeling. Construct validity is established by examining constructs (i.e., latent variables) that are not operationally defined or measured directly (Cronbach & Meehl, 1955). Confirmatory factor analysis allows the researcher to test specific hypotheses about how the measure of a construct is related to other measures based on theory or empirical evidence, or to examine constructs and their relations with other constructs. As has been already established, the six traits that are the focus of this study have been strongly established as a definition of writing and as measures for writing from both theoretical and empirical bases. The exploratory factor analyses conducted here, along with other statistical evidence, have shown that in the case of persuasive writing, it may be possible to reduce the six traits to three factors. Based on the definitions of the six traits and the semantic composition of the three factors, these three factors describe unique characteristics of writing. Confirmatory factor analyses can lend statistical support to the empirical evidence that has been mustered in support of the three-factor model and the theoretical evidence that has been used to further describe the three factors.

The analytic strategy I followed was to confirm the three-factor model on the 2004 DWA data and then replicate the three-factor model for each of the remaining years to add further validation of the model. If the three-factor model is not confirmed for the 2004 data, then I would resort to another exploratory factor analysis to modify the model and then confirm the modified model with the remaining years of data. AMOS 18 was used for all the analyses.

The proposed model based on the exploratory factor analysis appears in Figure 1. The three latent variables are patterned ideas, writer's perspective, and fluency components. The six-trait observed variables for patterned ideas are ideas and organization, each with an error term. The six-trait observed variables for writer's perspective are voice and word choice, with corresponding error terms.

Finally, the six-trait observed variables for fluency components are conventions and sentence fluency, with corresponding error terms. All three latent variables are proposed to be correlated with one another. The model was recursive with a sample size of 34,108. There were no missing data for these cases. There were 21 distinct sample moments, with 15 parameters to be estimated. Therefore, the model was over-specified with 6 df. Chi squared was 933.778, *p* , .001. The goodness-of-fit statistics are presented in Table 17, and the unstandardized and standardized parameter estimates are presented in Figures 2 and 3, respectively.

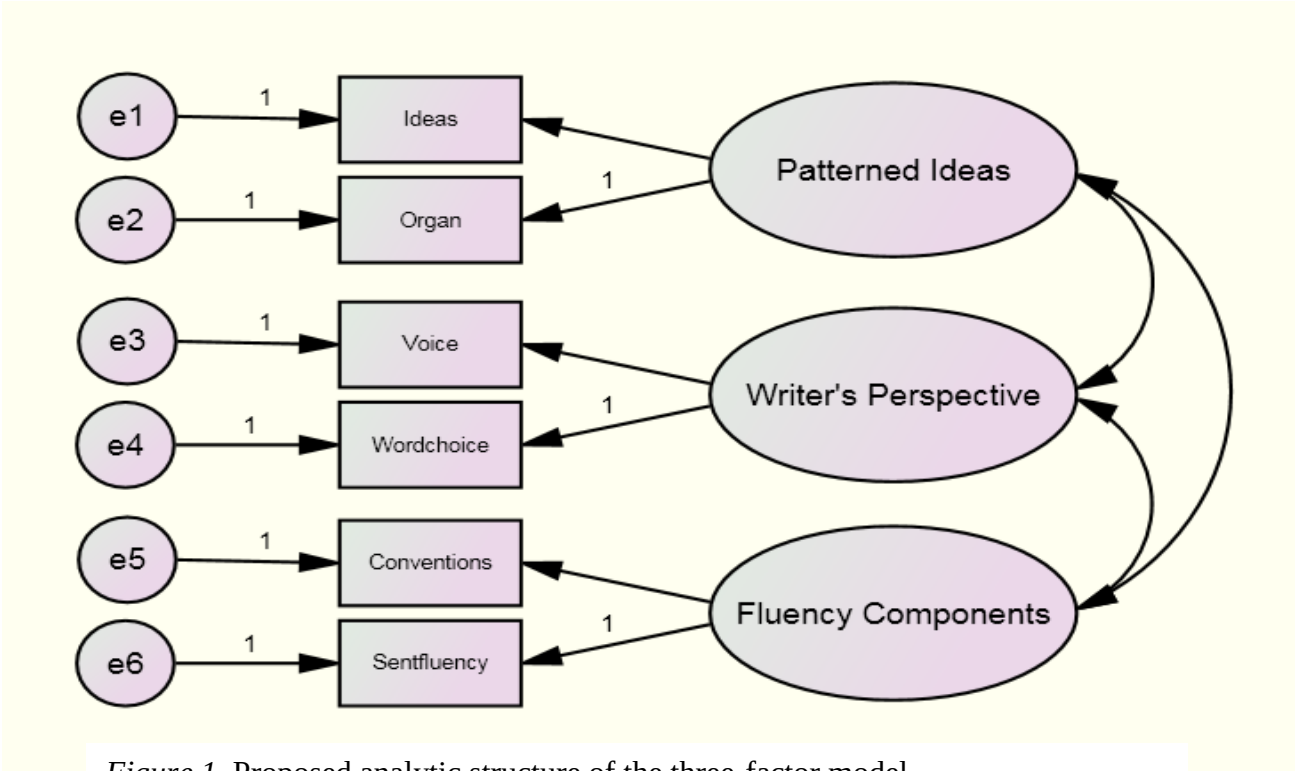

Figure 1. Proposed analytic structure of the three-factor model.

Table 17

Goodness-of-Fit Statistics for Proposed Three-Factor Model

RMR	GFI	CFI	RMSEA	Lo	Hi
.006	.991	.995	.067	.064	.071

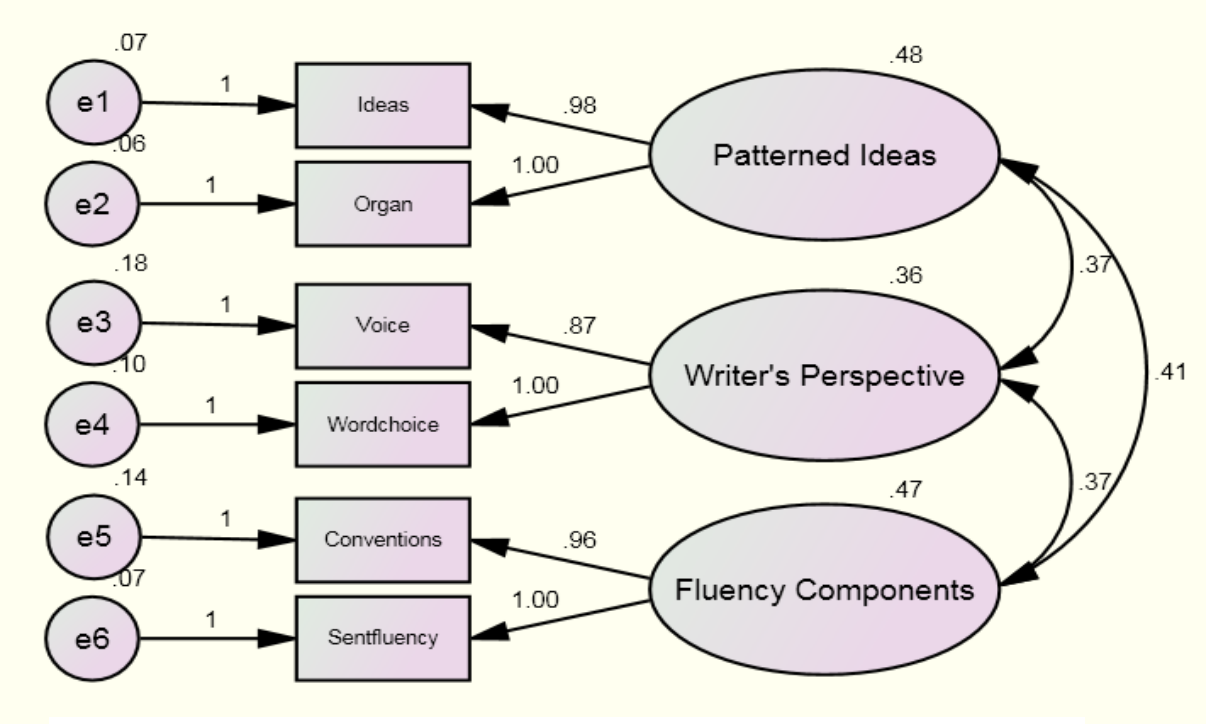

Figure 2. Unstandardized parameter estimates for the three-factor model.

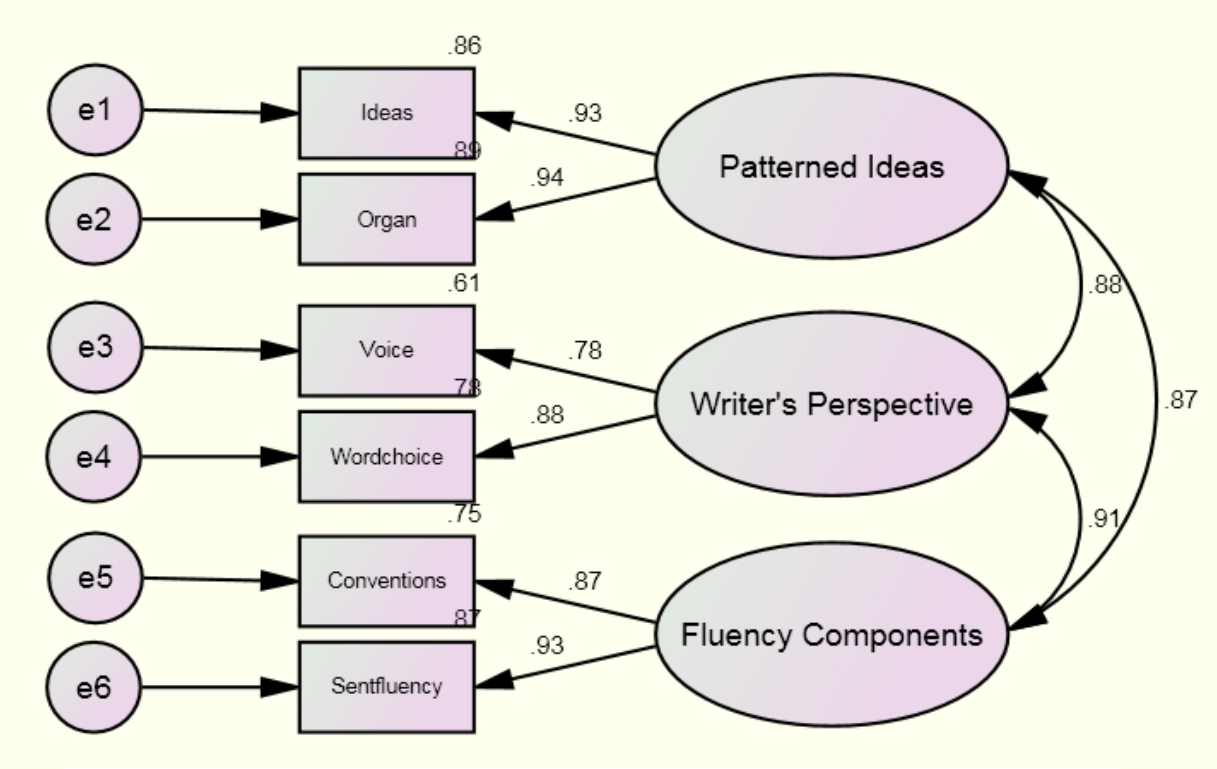

Figure 3. Standardized parameter estimates for the three-factor model.

The Root Mean Square Residual (RMR) of .006 was less than the recommended .05 for good model fit, and the Goodness of Fit Index (GFI) and Comparative Fit Index (CFI) were above .95, the recommendations for good model fit. Although the confidence interval is narrowly described, the Root Mean Square Error Approximation of .067 was high indicating a mediocre fit. Therefore, I modified the model using the modification indices, which indicated that adding a path from patterned ideas to voice could improve fit. The unstandardized and standardized models are presented in Figures 4 and 5, respectively, and the goodness-of-fit statistics are presented in Table 18.

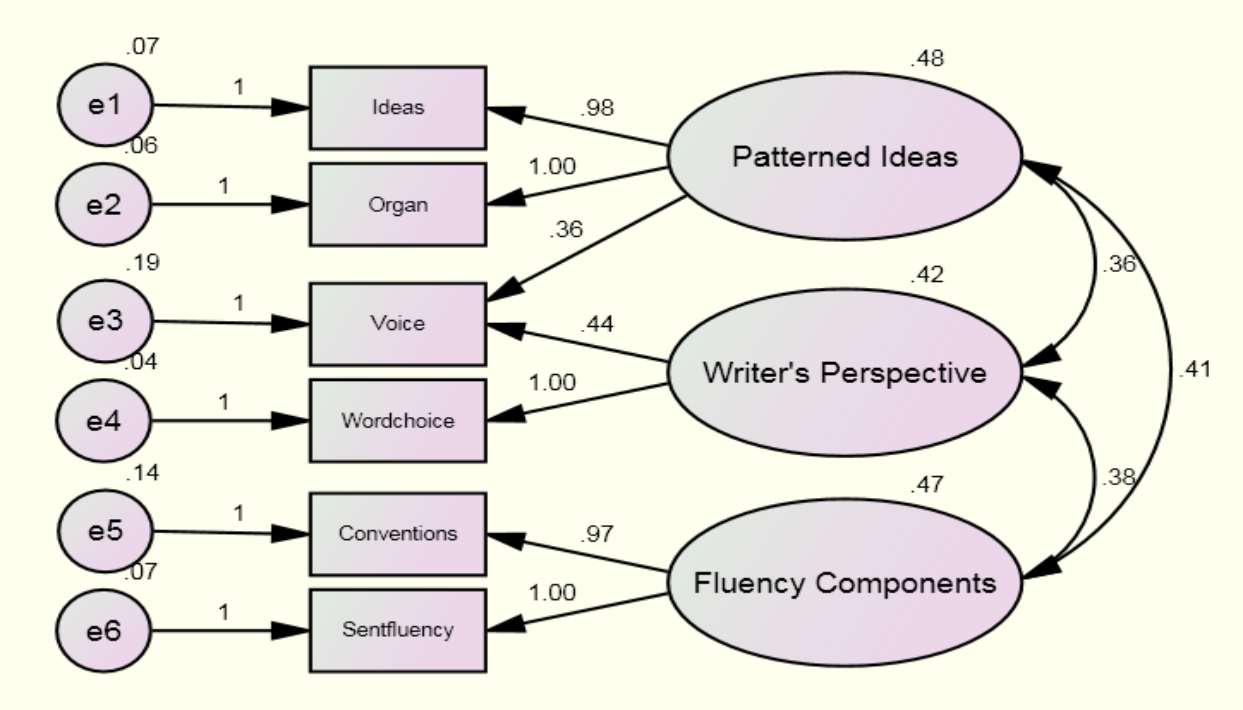

Figure 4. Unstandardized parameter estimates for the modified three-factor model.

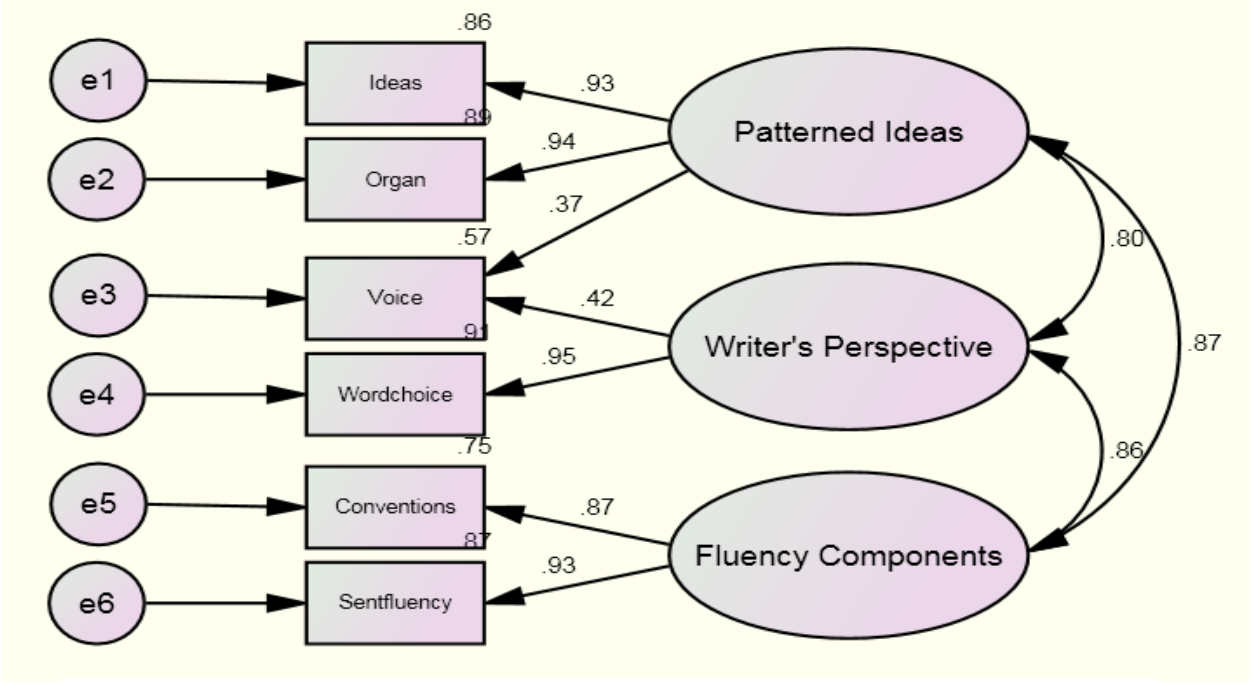

Figure 5. Standardized parameter estimates for the modified three-factor model.

Table 18

Goodness-of-Fit Statistics for Modified Three-Factor Model

RMR	GFI	CFI	RMSEA	Lo	Hi
.026	.998	.999	.030	.026	.034

All goodness-of-fit indices represent a good fit for the modified three-factor model. The regression weights between the two models stay approximately the same, with an expected drop between voice and writer's perspective with the additional path of .37 between voice and patterned ideas. Covariances and correlations among the factors remained approximately the same, and variances for factors and error terms remained approximately the same. Finally, a test of the Chi Squared values between the two models indicated a significant improvement: Original Model Chi Squared = 933.778; Modified Model Chi Squared = 156.885; 933.778 – 156.885 = 776.893 (1 df), $p < .001$.

The added path indicates that there is a moderate relation between voice and the latent factor of patterned ideas. "Patterned ideas" is defined as ideas within an essay and how those ideas are structured. Because the genre used for all 4 years of the data was persuasive writing, an argument can be made that the ideas and their pattern will reflect voice, which is defined as the unique perspective of the writer. Therefore, the inclusion of voice with patterned ideas makes sense.

Because the original three-factor model required modification, the confirmatory factor analysis became an exploratory factor analysis that requires further confirmation. To this end, I used data from 2005, 2006, and 2007 to validate this modified three-factor model. The goodness-of-fit statistics for each year are presented in Table 19. For each

Table 19

Goodness-of-Fit Statistics for Modified Three-Factor Model for Years 2005, 2006, and 2007

Year	RMR	GFI	CFI	RMSEA	Lo	Hi
2005	.003	.996	.998	.048	.044	.053
2006	.002	.997	.999	.042	.038	.046
2007	.002	.996	.998	.051	.047	.055

year, all goodness-of-fit indices represent a good fit for the modified three-factor model. Moreover, regression weights, variances, covariances, and correlations differ only slightly across the three years. The models with standardized estimates are presented in Figures 6, 7, and 8. This presents strong support for the validity of this model.

In sum, the internal structure of the DWA was validated across the 4 years of data. The three-factor model represents an adequate description of writing as it measured by the six traits.

I also validated the three-factor model using the various subgroups defined by ethnicity and ELL status. The same confirmatory factor analyses were conducted. Members from each group, based on either ethnicity or ELL status, were combined across the 4 years of data to provide sufficiently large samples. The summary of the analyses for ethnicity is presented in Table 20. Similar results were found for ELL status. For each ethnic group, the internal structure as represented by the three-factor model represents an adequate description of writing as it is measured by the six traits.

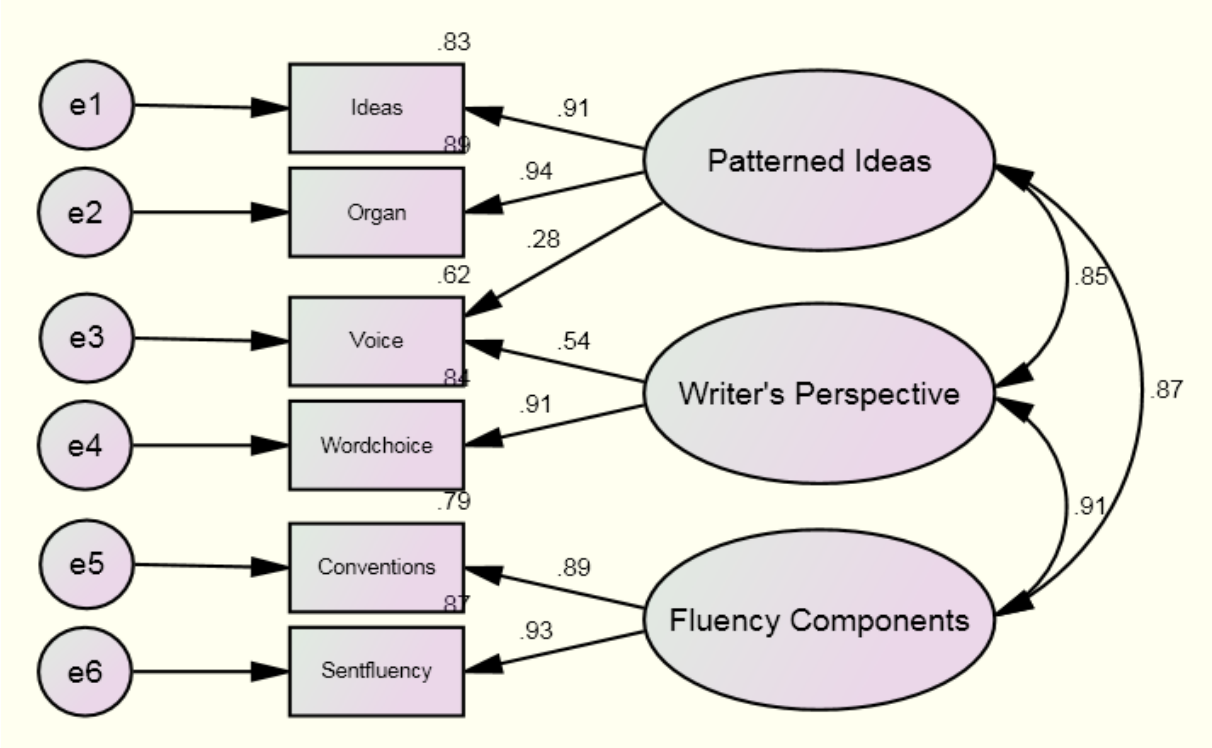

Figure 6. Standardized parameter estimates for the modified three-factor model for 2005.

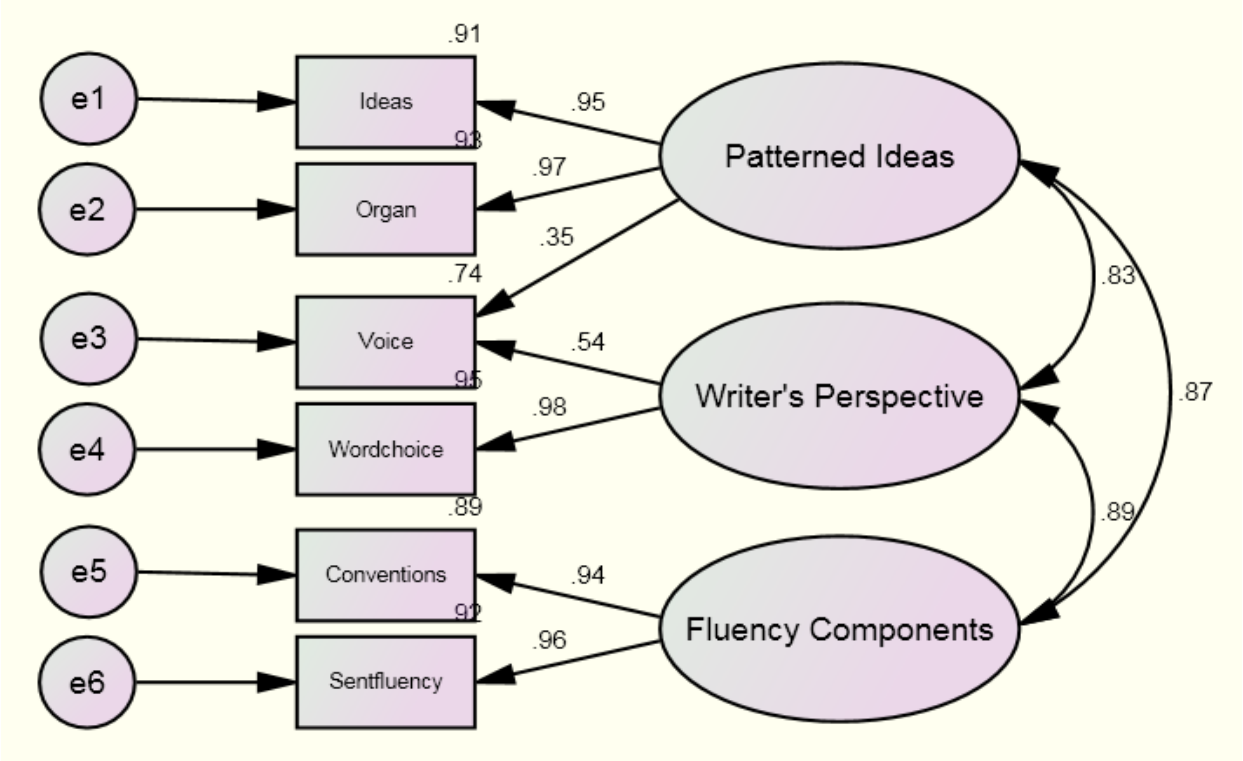

Figure 7. Standardized parameter estimates for the modified three-factor model for 2006.

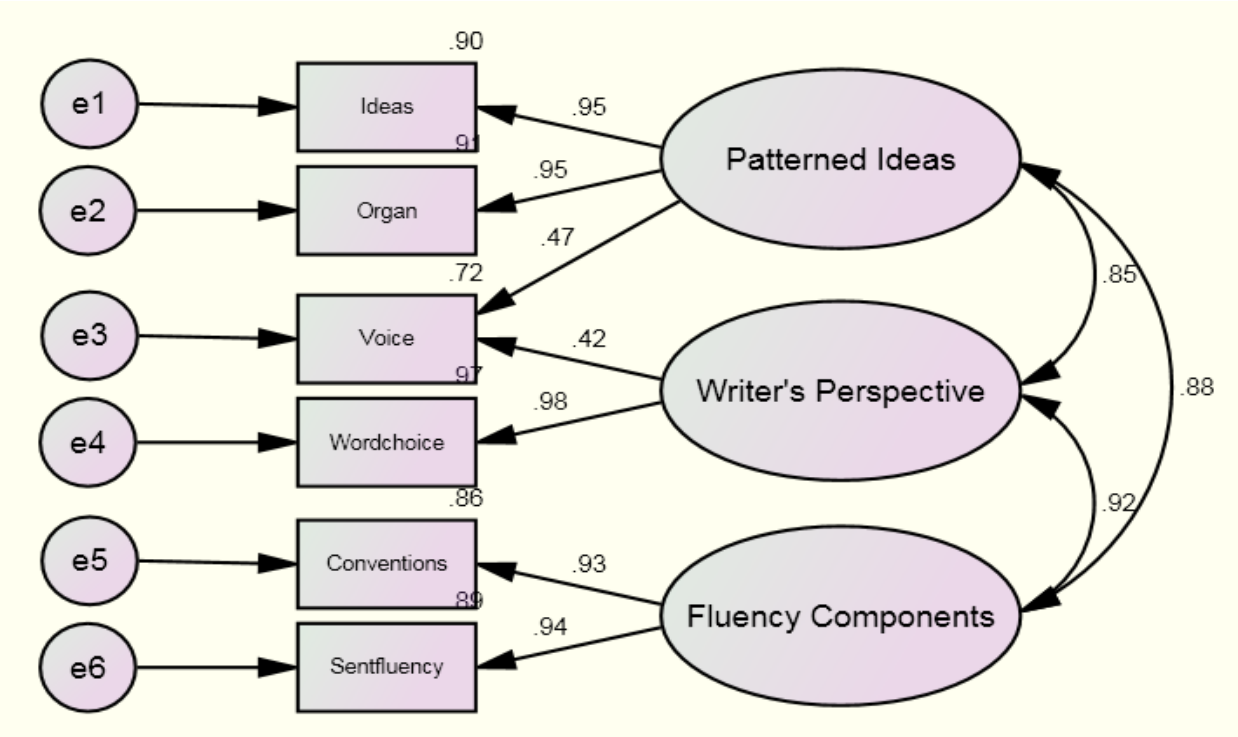

Figure 8. Standardized parameter estimates for the modified three-factor model for 2007.

Table 20

Goodness-of-Fit Statistics for Modified Three-Factor Model Tested for Ethnicity

Ethnicity	*n*	RMR	GFI	CFI	RMSEA	Lo		Hi	
Asian	2,206	.002	.997	.999	.036	.020		.053	
African-American	1,375	.004	.992	.997	.065	.046		.087	
Hispanic	13,198	.002	.998	.999	.034	.028	.041		
American Indian	2,212	.003	.994	.998	.054	.039	.071		
Pacific Islander	1,785	.002	.997	.999	.038	.020	.058		

Research Questions 2, 3, and 4

2) What is the relation between ELL status and writing proficiency for ninth-grade students attending public schools in Utah during the years 2004, 2005, 2006, and 2007, and to what extent do student variables, gender, social economic status, and ethnicity, independently and cumulatively explain the relation?

3) To what extent do the school variables, percent low-income students in a school, percent minority students in a school, size of the school, and mean ELL status at a school independently and cumulatively explain the relation between ELL status and writing proficiency?

4) To what extent do the district variables, percent of low-income students in a district, percent minority students in a district, size of the district, mean ELL status in a district, and whether a district is urban or rural independently and cumulatively explain the relation between ELL status and writing proficiency?

Answers to these three research questions were generated for each of the 4 years of data, using a three-level hierarchical linear model, with student at Level 1, school at Level 2, and district at Level 3. HLM 6 (Raudenbush, Bryk, & Congdon, 2004) was used for the analyses. Student writing proficiency (i.e., total 6 trait score) was analyzed as a function of gender (dichotomous variable), English Language Learning (1 to 6 scale, centered), ethnicity (dichotomous variable measuring whether students were White or non-White), and SES (dichotomous variable measuring whether students receive or do not receive free or reduced lunch programs). At the school level, variation across schools was analyzed according to the size of the school (number of students in ninth grade), the

mean ELL status at ninth grade, the proportion of ninth-grade students who were non-White, and the proportion of ninth-grade students in the school who were low-income (receiving free or reduced lunch). Finally, at the district level, variation across districts was analyzed according to the size of the district (number of students in ninth grade), the district mean ELL status, the proportion of the ninth-grade students in the district who were non-White, the proportion of ninth-grade students in the district who were low-income, and whether the district is urban or rural.

2004 Data

Table 21 shows that substantial amounts of variability can be explained by differences among schools and districts. About 24% of the variability among schools within districts and 25% of the variability among districts can be explained by differences between males and females. Similarly, about 34% of the variability among schools within districts and 32% of the variability among districts can be explained by differences between students on free or programs and students who pay for lunch. About 24% of the

Table 21

Level 1 and Level 2 Reliability Estimates 2004

Level 1		*Level 2*	
Random Coefficient	*Reliability*	*Random Coefficient*	*Reliability*
Gender	.24	Gender	.25
Low-income	.34	Low-income	.32
White/Non-White	.24		
ELL	.28	ELL	.29

variability among schools within districts can be explained by differences between White and non-White students. Finally, about 28% of the variability among schools within districts and 29% of the variability among districts can be explained by differences among the various ELL statuses. These results are typically encountered in cross-sectional studies of school effects where 10% to 30% of the variability is among schools (Raudenbush & Bryk, 2002).

The final conditional model is represented by the following equations:

Level 1 Model: Individual

$$Y_{ijk} = \pi_{0jk} + \pi_{1jk}(\text{gender}) + \pi_{2jk}(\text{low-income}) + \pi_{3jk}(\text{White/non-White}) + \pi_{4jk}(\text{ELL-centered}) + \varepsilon_{ijk}$$

Level 2 Model: School

$$\pi_{0jk} = \beta_{00k} + \beta_{01k}(\text{proportion low-income}) + \beta_{02k}(\#\text{ of students}) + r_{0jk}$$

$$\pi_{1jk}\,(\text{gender}) = \beta_{10k} + r1_{jk}$$

$$\pi_{2jk}\,(\text{low-income}) = \beta_{20k} + r_{2jk}$$

$$\pi_{3jk}\,(\text{White/non-White}) = \beta_{30k} + r_{3jk}$$

$$\pi_{4jk}\,(\text{ELL-centered}) = \beta_{40k} + \beta_{41k}(\#\text{ of students}) + r_{4jk}$$

Level 3 Model: District

$$\beta_{00k} = \gamma_{000}$$

$$\beta_{01k} = \gamma_{010}$$

$$\beta_{02k} = \gamma_{020}$$

$$\beta_{10k} = \gamma_{100} + U_{10k}$$

$$\beta_{20k} = \gamma_{200} + U_{20k}$$

$$\beta_{30k} = \gamma_{300} + \gamma_{301}(\text{urban/rural})$$

$$\beta_{40k} = \gamma_{400} + U_{40k}$$

$$\beta_{41k} = \gamma_{410}$$

The results for year 2004 show that there were significant differences due to gender, income, ethnicity, and ELL classification. The analyses confirm my expectations that there were differences in writing proficiency based on ELL status: Greater proficiency in English was associated with greater writing proficiency. This relation was moderated, however, with the size of the ninth-grade population at a school. With an increase in the ninth-grade population at a school, the relation between ELL status and writing proficiency got stronger. This moderation effect can be illustrated by first splitting the schools into thirds based on ninth-grade population, then examining the Pearson correlations between ELL status and writing score at the top and bottom third groups. For the largest and smallest schools, the correlations are .51 and .37, respectively ($ps < .01$).

The effects of income were seen at the individual, school, and district levels. Across all districts, students who received financial assistance for lunch programs scored lower by about 1.24 points than students who paid for lunch. Also, the proportion of students at a school who received free or reduced lunch contributed independently to writing proficiency. As the proportion of students who received financial assistance for lunch increased, writing scores decreased. To further illustrate this effect of low-income,

the Pearson correlation between proportion of students receiving assistance for lunch at a school and mean writing score at a school was -.55.

The size of the ninth-grade population at a school also contributed independently to writing proficiency. Schools with larger ninth-grade populations tended to score higher than schools with smaller ninth-grade populations. This effect is illustrated in Table 22 and further illustrated by a Pearson correlation of .44 between size of ninth-grade population and writing score.

With Whites coded as 0 and non-Whites as 1, the negative coefficient indicates that as the proportion of non-Whites to Whites approached 0 across districts, writing scores increased. However, this relation was moderated by whether a district was urban or rural. Non-White students tend to score higher in urban districts than in rural districts.

Finally, as expected, females outperformed males. As indicated in Table 22, this difference across all districts was about 1.43 points.

2005 Data

The 2005 data differ from other years in that DWA writing scores were not reported for four rural districts (Daggett, Logan, Millar, and Wayne) and four urban districts (Murray, Odgen, Provo, and Salt Lake City). The reduction of the number of districts from 40 to 32 may explain the loss of any district effects in the analyses. Table 23 shows that substantial amounts of variability can be explained by differences among schools. About 24% of the variability among schools within districts

Table 22

Parameter Estimates, Standard Errors, T-Ratios, and p-Values for Fixed Effects 2004

Student Parameters	School Parameters	District Parameters	Variables	Coefficient	Standard Error	*T*-ratio	*p*-value
π_{0jk}	β_{00k}	γ_{000}	Intercept	21.48	.13	161.85	< .001
	β_{01k}	γ_{010}	Proportion low-income at school	-2.19	.51	-4.29	< .001
	β_{02k}	γ_{020}	# of students in school	.002	.0008	2.40	= .018
π_{1jk}	β_{10k}	γ_{100}	Gender	1.43	.06	23.07	< .001
π_{2jk}	β_{20k}	γ_{200}	Low income	-1.24	.09	-13.06	< .001
π_{3jk}	β_{30k}	γ_{300}	White/ non-White	-.69	.13	-5.38	< .001
		γ_{301}	Urban/Rural	.46	.16	2.96	= .004
π_{4jk}	β_{40k}	γ_{400}	ELL	.63	.07	9.13	< .001
	β_{41k}	γ_{410}	School size	.001	.0003	4.81	< .001

Table 23

Level 1 and Level 2 Reliability Estimates 2005

Level 1		*Level 2*	
Random Coefficient	*Reliability*	*Random Coefficient*	*Reliability*
Gender	.24		
Low-income	.22		
ELL	.36		

can be explained by differences between males and females. About 22% of the variability among schools within districts can be explained by differences between students on free or reduced lunch programs and students who pay for lunch. Finally, about 36% of the variability among schools within districts can be explained by differences among the various ELL statuses.

The final conditional model is represented by the following equations:

Level 1 Model: Individual

$$Y_{ijk} = \pi_{0jk} + \pi_{1jk}(\text{gender}) + \pi_{2jk}(\text{low-income}) + \pi_{3jk}(\text{ELL-centered}) + \varepsilon_{ijk}$$

Level 2 Model: School

$$\pi_{0jk} = \beta_{00k} + \beta_{01k}(\text{proportion low-income}) + \beta_{02k}(\#\text{ of students}) + r_{0jk}$$

$$\pi_{1jk}\,(\text{gender}) = \beta_{10k} + \beta_{11k}(\text{proportion low-income}) + r1_{jk}$$

$$\pi_{2jk}\,(\text{low-income}) = \beta_{20k} + \beta_{21k}(\text{proportion low-income}) + r_{2jk}$$

$$\pi_{3jk} = \beta_{30k} + r_{3jk}$$

Level 3 Model: District

$$\beta_{00k} = \gamma_{000} + U_{00k}$$

$$\beta_{01k} = \gamma_{010}$$

$$\beta_{02k} = \gamma_{020}$$

$$\beta_{10k} = \gamma_{100}$$

$$\beta_{11k} = \gamma_{110}$$

$$\beta_{20k} = \gamma_{200}$$

$$\beta_{21k} = \gamma_{210}$$

$$\beta_{30k} = \gamma_{300}$$

The results for the year 2005 show that there were significant differences due to gender, income, and ELL classification. Ethnicity, however, showed no significant effects. Once again, as expected, ELL status contributed significantly to writing proficiency, with greater proficiency in English being associated with greater writing proficiency. There were no moderating effects on ELL status, but there were for the 2004 data.

The effects of income were seen at both the individual and school levels, but not the district level. Across all districts, students who received financial assistance for lunch programs scored lower by about 1.30 points than students who paid for lunch. This was similar to 2004, wherein this difference was 1.24. Also, as with the 2004 data, the proportion of students at a school who received financial assistance for lunch contributed independently to writing proficiency. As the proportion of students who received financial assistance for lunch increased, writing scores decreased. To further illustrate this effect of income, the Pearson correlation between proportion of students receiving assistance for lunch at a school and mean writing score at a school was -.64.

The proportion of students who received assistance with lunch programs also moderated the effect of gender, which, as expected, was that females outperformed males. The positive value of the coefficient indicates that as the proportion of low-income students at a school increases, the relation between gender and writing scores increases. This moderation effect can be illustrated by first splitting the schools into thirds based on the proportion of students receiving financial assistance for lunch programs and then

examining the differences between male and female writing scores (see Table 24). For schools with the smallest proportion of students receiving financial assistance for lunch programs, the difference between the mean scores for males and females was 1.39; for schools with the largest proportion, this increased to 1.72.

The proportion of students at a school who received assistance with lunch programs also moderated the effect of income on writing proficiency. The positive coefficient of 1.03 indicates that as the proportion of students at a school who receive free or reduced lunch increases, the effect of income on writing proficiency increases. That is, low-income students in schools with higher proportions of low-income students have lower writing skills than low-income students in schools with lower proportions of low-income students. This moderation effect is again illustrated by first splitting the schools into thirds based on the proportion of students receiving financial assistance for lunch programs and then examining the differences between the means for low-income students. For low-income students from schools with the smallest proportion of students on free or reduced lunch, their mean writing score was 20.80, and for low-income students from schools with the largest proportion of students on free or reduced lunch, their mean writing score was 19.46.

Finally, the size of the ninth-grade population at a school again contributed independently to writing proficiency. The positive value on the coefficient indicates that schools with larger ninth-grade populations had higher writing scores than schools with smaller ninth-grade populations. This relation is further illustrated by a Pearson correlation of .41 between size of ninth-grade population and writing score.

Table 24

Parameter Estimates, Standard Errors, T-Ratios, and p-Values for Fixed Effects 2005

Student Parameters	School Parameters	District Parameters	Variables	Coefficient	Standard Error	*T*-ratio	*p*-value
π_{0jk}	β_{00k}	γ_{000}	Intercept	21.47	.29	74.94	< .001
	β_{01k}	γ_{010}	Proportion Low-income at school	-3.19	.75	-4.26	< .001
	β_{02k}	γ_{020}	# of students in school	.002	.0009	2.62	= .010
π_{1jk}	β_{10k}	γ_{100}	Gender	1.29	.11	11.91	< .001
	β_{11k}	γ_{110}	Proportion Low Income at school	.86	.37	2.62	= .022
π_{2jk}	β_{20k}	γ_{200}	Low Income for student	-1.30	.15	-8.87	< .001
	β_{21k}	γ_{210}	Proportion Low Income at school	1.03	.44	2.36	= .020
π_{3jk}	β_{30k}	γ_{300}	ELL	.84	.06	14.89	< .001

2006 Data

Table 25 shows that substantial amounts of variability can be explained by differences among schools and districts. About 25% of the variability among schools within districts can be explained by differences between males and females. About 22% of the variability among schools within districts and 9% of the variability about districts can be explained by differences between students on free or reduced lunch programs and students who pay for lunch. The effect is illustrated in Table 26. Twenty-eight percent of the variability among schools within districts can be explained by differences between Whites and non-Whites. Finally, 36% of the variability among schools within districts can be explained by differences among the various ELL statuses.

Table 25

Level 1 and Level 2 Reliability Estimates 2006

Level 1		*Level 2*	
Random Coefficient	*Reliability*	*Random Coefficient*	*Reliability*
Gender	.25		
Low-income	.22	Low-income	.09
White/Non-White	.28		
ELL	.36		

Table 26

Parameter Estimates, Standard Errors, T-Ratios, and p-Values for Fixed Effects 2006

Student Parameters	School Parameters	District Parameters	Variables	Coefficient	Standard Error	*T*-ratio	*p*-value
π_{0jk}	β_{00k}	γ_{000}	Intercept	21.05	.16	129.44	< .001
	β_{01k}	γ_{010}	Proportion non-White at school	-2.60	.70	-3.72	< .001
	β_{02k}	γ_{020}	# of students in school	.003	.001	2.48	= .014
π_{1jk}	β_{10k}	γ_{100}	Gender	1.53	.05	29.85	< .001
	β_{11k}	γ_{110}	# of students in school	-.001	.0003	-3.27	< .01
π_{2jk}	β_{20k}	γ_{200}	Low income	-1.05	.07	-14.52	< .001
π_{3jk}	β_{30k}	γ_{300}	White/ non-White	-.68	.13	-5.20	< .001
		γ_{301}	# of schools in district	.05	.009	5.90	< .001
π_{4jk}	β_{40k}	γ_{400}	ELL	1.04	.11	9.19	< .001
		γ_{401}	Proportion low-income in district	-1.64	.31	-5.26	< .001
	β_{41k}	γ_{410}	# of students in school	.0008	.0004	2.37	= .019

The final conditional model is represented by the following equations:

Level 1 Model: Individual

$$Y_{ijk} = \pi_{0jk} + \pi_{1jk}(\text{gender}) + \pi_{2jk}(\text{low-income}) + \pi_{3jk}(\text{white/non-white}) + \pi_{4jk}(\text{ELL-centered}) + \varepsilon_{ijk}$$

Level 2 Model: School

$$\pi_{0jk} = \beta_{00k} + \beta_{01k}(\text{proportion non-White}) + \beta_{02k}(\text{\# of students}) + r_{0jk}$$

$$\pi_{1jk}(\text{gender}) = \beta_{10k} + \beta_{11k}(\text{\# of students}) + r1_{jk}$$

$$\pi_{2jk}(\text{low-income}) = \beta_{20k} + r_{2jk}$$

$$\pi_{3jk}(\text{White/non-White}) = \beta_{30k} + r_{3jk}$$

$$\pi_{4jk}(\text{ELL-centered}) = \beta_{40k} + \beta_{41k}(\text{\# of students}) + r_{4jk}$$

Level 3 Model: District

$$\beta_{00k} = \gamma_{000}$$

$$\beta_{01k} = \gamma_{010}$$

$$\beta_{02k} = \gamma_{020}$$

$$\beta_{10k} = \gamma_{100}$$

$$\beta_{11k} = \gamma_{110}$$

$$\beta_{20k} = \gamma_{200} + U_{20k}$$

$$\beta_{30k} = \gamma_{300} + \gamma_{301}(\text{\# of schools})$$

$$\beta_{40k} = \gamma_{400} + \gamma_{401}(\text{proportion low-income})$$

$$\beta_{41k} = \gamma_{410}$$

The results for year 2006 show that there were significant differences due to gender, income, ethnicity, and ELL classification. Once again, differences by ELL status indicate that greater proficiency in English was associated with greater writing proficiency. This relation was moderated independently by the size of the ninth-grade population at a school and by the proportion of students who received financial assistance for lunch programs at the district level. As with the 2004 data, the moderation effect for size of ninth-grade population was small but significant. There was a stronger relation between ELL status and writing proficiency in schools with larger ninth-grade populations than in schools with smaller ninth-grade populations. For the moderation effect from the proportion of students receiving free or reduced lunch, the negative coefficient indicates that as the proportion of low-income students in a district increases, the relation between ELL status and writing proficiency decreases. Therefore, the effects of ELL status on writing proficiency appear to diminish in smaller schools and in districts with larger proportions of students from low-income families.

The proportion of non-White students at a school and the size of the ninth-grade population at a school both contributed independently to writing proficiency. The negative coefficient on proportion of non-White students at a school indicates that writing proficiency decreased with the higher the proportions of non-White students at a school. Similar to the 2004 and 2005 data, the positive coefficient on number of ninth-grade students at a school indicates that as ninth-grade populations increase, so does writing proficiency.

As with the previous 2 years, females outperformed males, in this case by about 1.54 points. However, unlike the relation between gender and writing proficiency for the

previous years, size of the ninth-grade population at a school moderated the relation. The negative coefficient indicates that as the ninth-grade population increased the relation between gender and writing decreased. That is, the effects of gender on writing were lessened in schools with larger ninth-grade populations.

Also similar to the previous 2 years, there were differences due to ethnicity. White students outperformed non-White students. This effect between ethnicity and writing proficiency was moderated by the number of schools in a district. The positive coefficient indicates that as the number of schools in a district increased, the relation between ethnicity and writing increased. That is, in larger districts, differences between White and non-White students increased.

Last, similar to the previous 2 years, there was an effect between students who received financial assistance for lunch programs and those who paid for lunch. Students who paid for lunch outperformed students who received financial assistance for lunch programs.

2007 Data

Table 27 shows that substantial amounts of variability can be explained by differences among schools and districts. About 41% of the variability among schools within districts can be explained by differences between males and females. About 34% of the variability among schools within districts can be explained by differences between students on free or reduced lunch and students who pay full price for lunch. Twenty-nine percent of the variability among schools within districts and about 1% among districts can be explained by differences between Whites and non-Whites. Finally, 42% of the

Table 27

Level 1 and Level 2 Reliability Estimates 2007

Level 1		*Level 2*	
Random Coefficient	*Reliability*	*Random Coefficient*	*Reliability*
Gender	.41		
Low-income	.34		
White/Non-White	.29	White/Non-White	.001
ELL	.42		

variability among schools within districts can be explained by differences among the various ELL statuses.

The final conditional model is represented by the following equations:

Level 1 Model: Individual

$$Y_{ijk} = \pi_{0jk} + \pi_{1jk}(\text{gender}) + \pi_{2jk}(\text{low-income}) + \pi_{3jk}(\text{White/non-White}) + \pi_{4jk}(\text{ELL-centered}) + \varepsilon_{ijk}$$

Level 2 Model: School

$$\pi_{0jk} = \beta_{00k} + \beta_{01k}(\text{proportion low-income}) + \beta_{02k}(\#\text{ of students}) + r_{0jk}$$

$$\pi_{1jk}(\text{gender}) = \beta_{10k} + \beta_{11k}(\text{proportion low-income}) + r1_{jk}$$

$$\pi_{2jk}(\text{low-income}) = \beta_{20k} + r_{2jk}$$

$$\pi_{3jk}(\text{White/non-White}) = \beta_{30k} + r_{3jk}$$

$$\pi_{4jk}(\text{ELL-centered}) = \beta_{40k} + \beta_{41k}(\text{mean ELL})) + \beta_{42k}(\#\text{ of students}) + r_{4jk}$$

Level 3 Model: District

$$\beta_{00k} = \gamma_{000}$$

$$\beta_{01k} = \gamma_{010}$$

$$\beta_{02k} = \gamma_{020}$$

$$\beta_{10k} = \gamma_{100}$$

$$\beta_{11k} = \gamma_{110}$$

$$\beta_{20k} = \gamma_{200}$$

$$\beta_{30k} = \gamma_{300} + \gamma_{301}(\text{\# of schools}) + U_{30k}$$

$$\beta_{40k} = \gamma_{400}$$

$$\beta_{41k} = \gamma_{410}$$

$$\beta_{42k} = \gamma_{420}$$

The results for year 2007 show that there were significant differences due to gender, income, ethnicity, and ELL classification. As with the previous years, differences by ELL status indicate that greater proficiency in English was associated with greater writing proficiency. This effect was moderated independently by the size of the ninth-grade population at a school and by the mean ELL status in a school. The effect is illustrated in Table 28. There was a stronger relation between ELL status and writing proficiency in schools with larger ninth- grade populations than in schools with smaller ninth-grade populations. Separating schools into thirds based on ninth-grade population

Table 28

Parameter Estimates, Standard Errors, T-Ratios, and p-Values for Fixed Effects 2007

Student Parameters	School Parameters	District Parameters	Variables	Coefficient	Standard Error	*T*-ratio	*p*-value
π_{0jk}	β_{00k}	γ_{000}	Intercept	21.51	.32	67.96	< .001
	β_{01k}	γ_{010}	Proportion low-income at school	-2.70	.86	-3.14	< .01
	β_{02k}	γ_{020}	# of students in school	.002	.0008	2.29	= .023
π_{1jk}	β_{10k}	γ_{100}	Gender	1.19	.13	9.18	< .001
	β_{11k}	γ_{110}	Proportion low-income at school	.99	.40	2.50	= .014
π_{2jk}	β_{20k}	γ_{200}	Low income	-.98	.07	-13.85	< .001
π_{3jk}	β_{30k}	γ_{300}	White/ non-white	-.50	.11	-4.61	< .001
		γ_{301}	# of schools in district	.04	.007	5.58	< .001
π_{4jk}	β_{40k}	γ_{400}	ELL	.53	.08	6.91	< .001
	β_{41k}	γ_{410}	Mean ELL in school	-.36	.13	-2.79	= .006
	β_{42k}	γ_{420}	# of students in school	.0008	.0004	2.21	= .028

and correlating ELL status and writing score shows that the Pearson correlations for the smallest and largest ninth-grade populations were .32 and .46, respectively.

The effect of ELL status also was moderated by the mean ELL status at a school. The negative coefficient indicates that with greater English proficiency of a ninth-grade population, the relation between ELL status and writing score decrease. Separating schools into thirds on the basis of mean ELL status and then correlating ELL status and writing score shows that the Pearson correlations for the highest and lowest ELL means were -.21 (not significant) and .42. Therefore, the effects of ELL status on writing proficiency are diminished in schools with smaller ninth-grade populations and in schools in which there is greater proficiency in English for ELL students.

Both low-income and size of the ninth-grade population at a school independently contribute to writing proficiency. Low-income has an effect on writing proficiency similar to other years of the data. Both the main effect of low-income and this contributing effect indicate that the higher the proportion of students who receive financial assistance at a school, the lower writing proficiency becomes. Once again, as the size of the ninth-grade population at a school increases, so does the writing proficiency of the students.

Once again, the effect of gender was observed, with females more proficient at writing than males. However, for this year of data, the effect of gender was moderated by the proportion of students who receive financial assistance for lunch programs. The positive coefficient indicates that as the proportion of students who receive financial assistance increases, so does the relation between gender and writing proficiency. That is, in schools with higher proportions of low-income students, the gap between males and

females increases. To illustrate this relation, schools first were split into thirds on the basis of proportion low-income, gender, and writing scores, which were correlated. For the schools with the lowest proportion of students on free or reduced lunch the Pearson correlation was .17 (not significant), and for the school with the highest proportion of students on free or reduced lunch the Pearson correlation was .28.

Finally, the same differences based on ethnicity were also observed: Whites scored higher than non-Whites. However, for this year, this effect was moderated by the number of schools in a district. The positive coefficient indicates that as the number of schools in a district increases, the relation between ethnicity and writing proficiency increases. That is, larger school districts see larger differences between Whites and non-Whites in writing proficiency.

Research Question 5

My fifth research question was: To what extent does each of the six traits vary with ELL status and ethnicity, and are there interactions between ELL status and ethnicity on any of the six traits? To answer this question, I conducted a multivariate analysis of variance (MANOVA), using the two categorical variables ELL status and ethnicity as independent variables and the six trait scores as continuous dependent variables. Thus, the analysis was an ELL status (5 levels) x Ethnicity (6 levels) MANOVA with six dependent variables. I tested for main effects for ELL status and ethnicity and for interactions between ELL status and ethnicity. Data from the 4 years were collapsed to provide sufficient number of individuals in each of the 30 cells. With the exception of three cells, the number of individuals in each cell exceeded 23.

American Indian—pre-emergent had no cases, Pacific Islander—pre-emergent had 9 cases, and African-American—pre-emergent had 14 cases. Therefore, with the deletion of the American Indian—pre-emergent cell, there are more cases than DVs in each cell. As reported earlier, all DVs were normally distributed.

SPSS MANOVA was used for the analysis, which adjusts for unequal *n*'s. With the use of Wilks' Lambda, the combined dependent variables were significantly affected by both ethnicity, $F(30, 45{,}130) = 5.77$, $p < .001$, Wilks' Lambda = .98, $\eta^2 = .003$, and ELL status, $F(24, 39{,}359) = 31.08$, $p < .001$, Wilks' Lambda = .94, $\eta^2 = .02$. There also was a significant interaction between ethnicity and ELL status, $F(120, 65{,}210) = 1.61$, $p < .001$, Wilks' Lambda = .98, $\eta^2 = .003$.

Because the omnibus MANOVA showed a significant interaction, I further investigated the nature of the relations among the IVs and DVs. There is no priority ordering of the DVs. Therefore, univariate F tests were used to assess the DVs rather than stepdown analyses. For the ethnicity x ELL status interaction, univariate *F*-tests (20, 11,287 df) were conducted for each of the six dependent variables (see Table 29). All six dependent variables were significant.

The MANOVA showed a significant main effect for ELL status, with each of the six traits significant (all *p*s < .001). Therefore, to bring further clarity to the effects of the interaction between ELL status and ethnicity on the DVs, I first examined ELL status by each of the ethnic groups by conducting a MANOVA with ELL status as the independent

Table 29

Univariate Analyses of Variance of Dependent Variables for Interaction Between Ethnicity and ELL Status

Dependent Variable	Sum of Squares	Mean Square	*F*	*p*-value
Ideas	28.16	1.41	2.79	< .001
Organization	22.73	1.14	2.23	= .001
Voice	20.12	1.01	2.30	= .001
Word Choice	22.41	1.21	2.64	< .001
Sentence Fluency	30.60	1.53	3.07	.001
Conventions	31.57	1.58	3.07	< .001

variable and the six traits as the dependent variables. For the following analyses (see Table 30), the five ELL statuses were used (i.e., pre-emergent, emergent, intermediate, advanced, advanced, and fluent).

For Asians, the combined dependent variables were significantly affected by ELL status, $F(24, 3{,}556) = 14.08$, $p < .001$, Wilks' Lambda = .73, $\eta^2 = .08$. Univariate *F*-tests (4, 1,024 df) were conducted for each of the six dependent variables (see Table 30). All six dependent variables were significant. Multiple comparisons using Tukey's HSD showed that fluent students outperformed advanced students who outperformed intermediate students who outperformed emergent and pre-emergent on all six traits (see Table 31). There were no differences between the latter two groups.

Table 30

Univariate Analyses of Variance for ELL Status on Each Dependent Variable for Asians

Dependent Variable	Sum of Squares	Mean Square	*F*	*p*-value	Partial Eta Squared
Ideas	117.52	29.38	65.85	< .001	.20
Organization	112.17	30.54	66.62	< .001	.21
Voice	76.95	19.24	46.61	< .001	.15
Word Choice	110.62	27.65	64.88	< .001	.20
Sentence Fluency	145.94	36.48	76.16	<.001	.23
Conventions	157.32	39.33	82.75	< .001	.24

Table 31

Writing Performance on Each of the Six Traits for Each of the ELL Statuses for Asians

	Pre-Em *n* = 23		Emerg *n* = 131		Interm *n* = 71		Adv *n* = 298		Fluent *n* = 506
Ideas	2.83 (.98)	≈	2.90 (.75)	<	3.27 (.61)	<	3.63 (.67)	<	3.85 (.64)
Organization	2.78 (1.00)	≈	2.88 (.78)	<	3.30 (.59)	<	3.65 (.67)	<	3.85 (.64)
Voice	3.00 (.74)	≈	3.12 (.72)	<	3.44 (.60)	<	3.67 (.66)	<	3.88 (.61)
Word Choice	2.70 (.82)	≈	2.88 (.65)	<	3.25 (.67)	<	3.55 (.64)	<	3.79 (.65)
Sentence Fluency	2.52 (.95)	≈	2.82 (.71)	<	3.17 (.74)	<	3.57 (.69)	<	3.83 (.67)
Conventions	2.43 (.79)	≈	2.79 (.73)	<	3.23 (.69)	<	3.56 (.69)	<	3.85 (.67)

For African Americans, the combined dependent variables were significantly affected by ELL status, $F(24, 507) = 14.08$, $p < .001$, Wilks' Lambda = 3.13, $\eta^2 = .11$. Univariate *F*-tests (4, 150 df) were conducted for each of the six dependent variables (see Table 32). All six dependent variables were significant. Multiple comparisons using Tukey's HSD showed no differences between fluent, advanced, or intermediate students; however, intermediate students outperformed emergent on all but word choice, and emergent outperformed pre-emergent on all six traits (see Table 33).

For Whites, the combined dependent variables were significantly affected by ELL status, $F(24, 3{,}915) = 8.70$, $p < .001$, Wilks' Lambda = .83, $\eta^2 = 04$. Univariate *F*-tests (4, 1,127 df) were conducted for each of the six dependent variables (see Table 34). All six dependent variables were significant. Multiple comparisons using Tukey's HSD showed that fluent students outperformed advanced students who outperformed intermediate students (see Table 35).

Table 32

Univariate Analyses of Variance for ELL Status on Each Dependent Variable for African Americans

Dependent Variable	Sum of Squares	Mean Square	*F*	*p*-value	Partial Eta Squared
Ideas	29.11	7.28	12.53	< .001	.25
Organization	30.22	7.56	13.08	< .001	.26
Voice	22.29	5.57	13.21	< .001	.26
Word Choice	26.27	6.57	14.14	< .001	.27
Sentence Fluency	39.16	9.79	16.32	< .001	.30
Conventions	47.35	11.84	18.37	< .001	.33

Table 33

Writing Performance on Each of the Six Traits for Each of the ELL Statuses for African Americans

	Pre-Em *n* = 14		Emerg *n* = 60		Interm *n* = 19		Adv *n* = 32		Fluent *n* = 30
Ideas	1.93 (.92)	<	2.67 (.68)	<	3.32 (.88)	≈	3.31 (.90)	≈	3.27 (.58)
Organization	1.86 (.86)	<	2.58 (.70)	<	3.16 (.83)	≈	3.28 (.92)	≈	3.23 (.57)
Voice	2.29 (.61)	<	2.92 (.59)	<	3.37 (.76)	≈	3.53 (.67)	≈	3.47 (.68)
Word Choice	2.00 (.88)	<	2.67 (.65)	≈	3.10 (.74)	≈	3.41 (.71)	≈	3.20 (.55)
Sentence Fluency	1.86 (.77)	<	2.53 (.79)	<	3.21 (.79)	≈	3.44 (.80)	≈	3.30 (.70)
Conventions	1.57 (.65)	<	2.43 (.81)	<	3.05 (.85)	≈	3.34 (.83)	≈	3.30 (.79)

Table 34

Univariate Analyses of Variance for ELL Status on Each Dependent Variable for Whites

Dependent Variable	Sum of Squares	Mean Square	*F*	*p*-value	Partial Eta Squared
Ideas	73.73	18.43	35.87	< .001	.11
Organization	80.07	20.02	39.9	< .001	.12
Voice	66.07	16.52	35.53	< .001	.11
Word Choice	66.52	16.63	36.75	< .001	.12
Sentence Fluency	86.01	21.50	44.30	< .001	.14
Conventions	79.32	19.83	37.90	< .001	.12

Table 35

Writing Performance on Each of the Six Traits for Each of the ELL Statuses for Whites

	Pre-Em *n* = 30		Emerg *n* = 145		Interm *n* = 86		Adv *n* = 336		Fluent *n* = 535
Ideas	2.97 (.85)	≈	2.97 (.82)	≈	3.09 (.75)	<	3.48 (.71)	<	3.65 (.68)
Organization	2.83 (.75)	≈	2.96 (.80)	≈	3.07 (.73)	<	3.47 (.71)	<	3.65 (.68)
Voice	3.03 (.96)	≈	3.16 (.73)	≈	3.15 (.70)	<	3.54 (.66)	<	3.76 (.66)
Word Choice	2.80 (.80)	≈	3.01 (.71)	≈	3.12 (.68)	<	3.42 (.66)	<	3.62 (.66)
Sentence Fluency	2.77 (.90)	≈	3.00 (.78)	≈	3.07 (.75)	<	3.45 (.69)	<	3.69 (.66)
Conventions	2.67 (.84)	≈	2.96 (.85)	≈	3.03 (.71)	<	3.44 (.70)	<	3.61 (.69)

There were no differences among intermediate, emergent, or pre-emergent students.

For Hispanics, the combined dependent variables were significantly affected by ELL status, $F(24, 26{,}162) = 79.56$, $p < .001$, Wilks' Lambda = .78, $\eta^2 = 06$. Univariate *F*-tests (4, 7,504 df) were conducted for each of the six dependent variables (see Table 36). All six dependent variables were significant. Multiple comparisons using Tukey's HSD showed that from fluent students to pre-emergent students, each ELL group outperformed the ELL group below it (see Table 37).

Table 36

Univariate Analyses of Variance for ELL Status on Each Dependent Variable for Hispanics

Dependent Variable	Sum of Squares	Mean Square	*F*	*p*-value	Partial Eta Squared
Ideas	824.33	206.08	402.24	< .001	.18
Organization	877.83	219.46	422.17	< .001	.18
Voice	603.82	150.96	345.68	< .001	.16
Word Choice	570.05	142.51	338.56	< .001	.15
Sentence Fluency	801.05	200.26	395.22	< .001	.17
Conventions	904.84	226.21	433.36	< .001	.19

Table 37

Writing Performance on Each of the Six Traits for Each of the ELL Statuses for Hispanics

	Pre-Em *n* = 321		Emerg *n* = 1870		Interm *n* = 1115		Adv *n* = 2320		Fluent *n* = 1883
Ideas	2.38 (.86)	<	2.73 (.72)	<	3.02 (.70)	<	3.33 (.70)	<	3.48 (.72)
Organization	2.39 (.88)	<	2.68 (.72)	<	2.98 (.72)	<	3.32 (.71)	<	3.47 (.71)
Voice	2.68 (.87)	<	2.92 (.66)	<	3.18 (.64)	<	3.44 (.64)	<	3.59 (.66)
Word Choice	2.60 (.95)	<	2.78 (.67)	<	3.02 (.60)	<	3.28 (.61)	<	3.44 (.64)
Sentence Fluency	2.52 (.95)	<	2.68 (.73)	<	2.96 (.71)	<	3.29 (.68)	<	3.48 (.68)
Conventions	2.41 (.92)	<	2.61 (.75)	<	2.92 (.73)	<	3.26 (.68)	<	3.44 (.69)

For American Indians, the combined dependent variables were significantly affected by ELL status, $F(24, 2{,}715) = 7.57$, $p < .001$, Wilks' Lambda = .80, $\eta^2 = .06$. Univariate *F*-tests (4, 783 df) were conducted for each of the six dependent variables (see Table 38). All six dependent variables were significant. Multiple comparisons using Tukey's HSD showed that with the exception of voice, there were no differences between exited and advanced groups. There were significant differences between advanced and intermediate on all six traits. There were significant differences between intermediate and emergent with the exception of voice (see Table 39).

Table 38

Univariate Analyses of Variance for ELL Status on Each Dependent Variable for American Indians

Dependent Variable	Sum of Squares	Mean Square	*F*	*p*-value	Partial Eta Squared
Ideas	78.92	19.73	37.42	< .001	.16
Organization	79.92	19.98	37.73	< .001	.16
Voice	50.66	12.67	29.36	< .001	.13
Word Choice	48.46	12.12	27.30	< .001	.12
Sentence Fluency	65.56	16.39	33.51	< .001	.15
Conventions	77.37	19.59	38.03	< .001	.16

Table 39

Writing Performance on Each of the Six Traits for Each of the ELL Statuses for American Indians

	Pre-Em *n* = 0		Emerg *n* = 324		Interm *n* = 120		Adv *n* = 178		Fluent *n* = 162
Ideas			2.61 (.76)	≈	2.79 (.68)	<	3.20 (.75)	≈	3.34 (.66)
Organization			2.61 (.77)	<	2.82 (.67)	<	3.20 (.74)	≈	3.36 (.67)
Voice			2.88 (.71)	<	3.07 (.62)	<	3.29 (.64)	<	3.51 (.59)
Word Choice			2.73 (.72)	<	2.97 (.59)	<	3.17 (.66)	≈	3.33 (.60)
Sentence Fluency			2.66 (.73)	<	2.88 (.64)	<	3.18 (.74)	≈	3.35 (.63)
Conventions			2.61 (.74)	<	2.89 (.66)	<	3.18 (.76)	≈	3.38 (.67)

For Pacific Islanders, the combined dependent variables were significantly affected by ELL status, $F(24, 2{,}422) = 8.32$, $p < .001$, Wilks' Lambda = .78, $\eta^2 = 07$. Univariate *F*-tests (4, 699 df) were conducted for each of the six dependent variables (see Table 40). All six dependent variables were significant. Multiple comparisons usingTukey's HSD showed that the exited group outperformed the advanced group on all six traits, the advanced group outperformed the intermediate only on conventions, the intermediate group outperformed the emergent on all six traits, and there were no differences between the latter group and the pre-emergent group (see Table 41).

Table 40

Univariate Analyses of Variance for ELL Status on Each Dependent Variable for Pacific Islanders

Dependent Variable	Sum of Squares	Mean Square	*F*	*p*-value	Partial Eta Squared
Ideas	55.30	13.82	30.37	< .001	.15
Organization	68.75	17.19	37.96	< .001	.18
Voice	51.30	12.83	28.32	< .001	.14
Word Choice	50.93	12.73	32.02	< .001	.16
Sentence Fluency	78.36	19.59	42.82	< .001	.20
Conventions	75.57	18.89	42.93	< .001	.20

Table 41

Writing Performance on Each of the Six Traits for Each of the ELL Statuses for Pacific Islanders

	Pre-Em *n* = 9		Emerg *n* = 127		Interm *n* = 57		Adv *n* = 228		Fluent *n* = 283
Ideas	2.67 (1.00)	≈	2.86 (.70)	<	3.32 (.60)	≈	3.42 (.66)	<	3.62 (.68)
Organization	2.67 (1.00)	≈	2.78 (.73)	<	3.21 (.53)	≈	3.40 (.67)	<	3.62 (.66)
Voice	2.89 (1.05)	≈	3.03 (.67)	<	3.47 (.57)	≈	3.50 (.65)	<	3.77 (.70)
Word Choice	2.44 (1.13)	≈	2.88 (.60)	<	3.19 (.52)	≈	3.35 (.64)	<	3.58 (.64)
Sentence Fluency	2.33 (1.00)	≈	2.75 (.71)	<	3.23 (.54)	≈	3.38 (.67)	<	3.63 (.68)
Conventions	2.56 (1.13)	≈	2.76 (.70)	<	3.12 (.63)	<	3.38 (.64)	<	3.63 (.65)

As an additional way to bring clarity to the differences in performance on the six traits by ethnicity, I collapsed the data across the 4 years and conducted a mixed analysis of variance (ANOVA) with ethnicity the between-subjects variable and the six traits as a within-subjects variable. There was a significance between subjects effect, $F(5, 130{,}921) = 1074.39$, $p < .001$, $\eta^2 = .04$, a significant within subjects effect, (Greenhouse-Geisser) $F(4.11, 538{,}593) = 398.61$, $p < .001$, $\eta^2 = .003$, and a significant ethnicity x six traits interaction¸(Greenhouse-Geisser) $F(20.57, 538{,}593) = 29.12$, $p < .001$, $\eta^2 = .001$. For the between subjects effect, Tukey's HSD post hoc tests showed that Asians outperformed Whites who outperformed Pacific Islanders who outperformed African Americans who outperformed Hispanics who outperformed American Indians (see Figure 9).

The significant interaction was further investigated by conducting a repeated measure ANOVA for each ethnic group with the six traits as the repeated measure. The results of the ANOVAs and follow-up tests comparing the six traits are illustrated in Figure 1 and described in detail in Table 42.

Each group performed significantly better on voice than any other trait. Differences among the groups were most evident by the traits on which performance was the lowest: Asians scored the lowest on word choice, Whites scored the lowest mostly on conventions and word choice, Pacific Islanders scored lowest on word choice, African-Americans scored the lowest on conventions, Hispanics scored lowest mostly on conventions, and American Indians scored lowest mostly on conventions and organization.

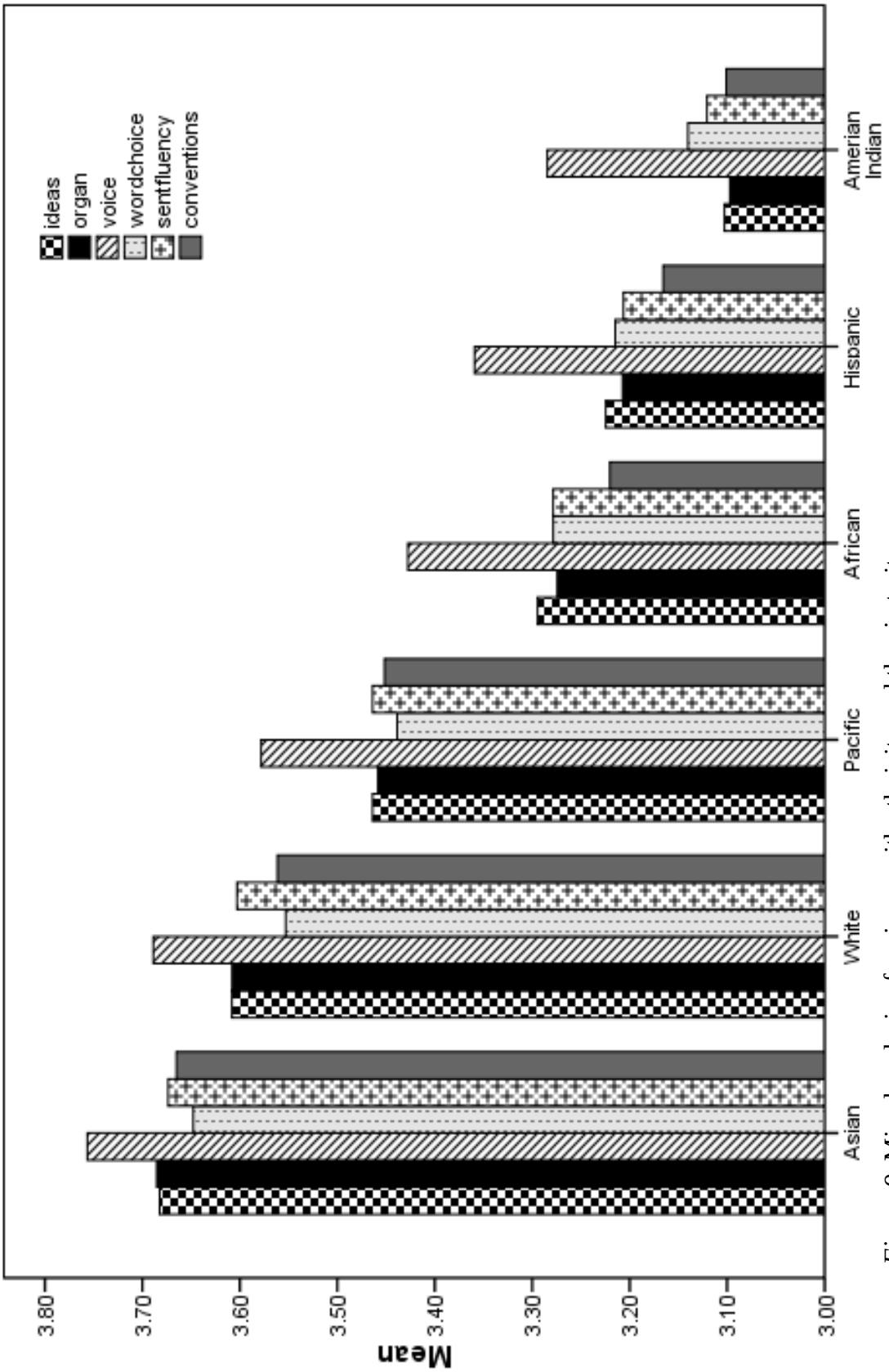

Figure 9. Mixed analysis of variance with ethnicity and the six traits.

Table 42

Differences Among Six Traits for Each Ethnic Group

Hispanics (n = 13,198) $F(5, 65{,}985) = 478.03$, $p < .001$, $\eta^2 = .035$				
Voice 3.36 (.72)	>	Ideas 3.22 (.79)	>	Word Choice Organization Sentence Fluency Conventions
	>	Word Choice 3.21 (.70)	>	Sentence Fluency Conventions
	>	Organization 3.21 (.79)	>	Conventions
	>	Sentence Fluency 3.21 (.77)	>	Conventions
	>	Conventions 3.17 (.79)		
Whites (n = 110,151) $F(5, 550{,}750) = 2{,}089.49$, $p < .001$, $\eta^2 = .019$				
Voice 3.69 (.67)	>	Ideas 3.61 (.71)	>	Sentence Fluency Conventions Word Choice
	>	Organization 3.61 (.72)	>	Sentence Fluency Conventions Word Choice
	>	Sentence Fluency 3.60 (.71)	>	Conventions Word Choice
	>	Conventions 3.56 (.73)	>	Word Choice
	>	Word Choice 3.55 (.68)		
Pacific Islanders (n = 1,785) $F(5, 8{,}920) = 39.53$, $p < .001$, $\eta^2 = .022$				
Voice 3.58 (.68)	>	Ideas 3.46 (.71)	>	Word Choice
	>	Sentence Fluency 3.46 (.73)	>	Word Choice
	>	Organization 3.46 (.72)		
	>	Conventions 3.45 (.72)		
	>	Word Choice 3.44 (.67)		

Table 42 Continued

American Indians (n = 2,212) $F(5, 11{,}025) = 90.66$, $p < .001$, $\eta^2 = .039$						
Voice 3.28 (.72)	>	Word Choice 3.14 (.73)	>	Sentence Fluency Ideas Conventions Organization		
	>	Sentence Fluency 3.12 (.78)	>	Conventions Organization		
	>	Ideas 3.10 (.81)				
	>	Conventions 3.10 (.79)				
	>	Organizations 3.10 (.81)				
African Americans (n = 1,375) $F(5, 6{,}870) = 51.33$, $p < .001$, $\eta^2 = .036$						
Voice 3.43 (.70)	>	Ideas 3.29 (.77)	>	Organization	>	Conventions
			>	Conventions		
	>	Word Choice 3.28 (.68)	>	Conventions		
	>	Sentence Fluency 3.28 (.77)	>			
	>	Organization 3.27 (.79)	>			
	>	Conventions 3.22 (.80)				
Asians (n = 2,206) $F(5, 11{,}025) = 26.25$, $p < .001$, $\eta^2 = .012$						
Voice 3.76 (.68)	>	Organization 3.68 (.74)	>	Word Choice		
	>	Ideas 3.68 (.73)	>	Word Choice		
	>	Sentence Fluency 3.67 (.75)	>	Word Choice		
	>	Conventions 3.66 (.76)				
	>	Word Choice 3.65 (.71)				

CHAPTER 5

DISCUSSION

With a clear sense that the majority of students in our country write below the writing proficiency standard (National Center for Educational Statistics, 2007), I was motivated to examine the writing skills of students, particularly ELL students, in the State of Utah, and identify variables that may contribute to those skills. Using 4 years (2004 – 2007) of ninth-grade DWA archival data, which was approximately a sample of 135,000 students, I had several goals in mind.

My initial goal was to examine the validity and reliability of the DWA scores for all students and for subgroups determined by ELL status and ethnicity. The DWA uses the six traits, so in a sense, I was investigating the validity and reliability of scores derived from six-trait scoring. The DWA must be providing valid scores that assess the psychological construct of writing skill, and those scores must be measured with acceptable precision. Second, I then assessed the relative contributions of student, school, and school district to the relationship between ELL status and writing proficiency. In addition, I identified predictor variables at the individual, school, and district level that

might contribute to this relationship. Some of the predictor variables chosen for this study are variables that have been previously linked to writing proficiency, such as students' gender, income, and ethnicity. By identifying variables that significantly contribute to the relation between ELL status and writing proficiency, I hoped to identify conditions that contribute to or detract from ELL students' writing proficiency. Once these conditions are identified, the groundwork is established for future empirical work that can more specifically investigate the school and classroom environments that promote writing skill. I ended with an examination of the differences among ethnic groups and ELL statuses in writing proficiency. Examining these differences might provide us with insights into cultural differences in writing practices and differences due to ELL status. Understanding these differences could inform future writing curriculum policy decisions.

Research Question 1

Does the DWA provide valid and reliable scores of writing proficiency for students in general and for specific groups of students based on ELL status and ethnicity?

"Validity refers to the degree to which evidence and theory support the interpretation of test scores entailed by proposed uses of tests" and is "the most fundamental consideration in developing and evaluating tests" (AERA, 1999, p. 9). Put another way, establishing validity for the uses of a test requires that theoretical and empirical evidence from multiple sources is used in building a case to "support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment" (Messick, 1989, p. 13). The various types of validity that have

been identified (i.e., construct, content, external, internal) serve as a way to organize empirical and theoretical evidence to support decisions based on a test's score (Osterlind, 2006). To establish the reliability—a necessary but insufficient criterion for validity—and validity of the DWA, I gathered multiple sources of theoretical and empirical evidence to build a case that the test scores from the DWA allow appropriate inferences and decisions.

I followed Osterlind's (2006) five guidelines for examining whole tests. First, I reviewed the published information about the DWA that was provided by the company that scores the test. Second, for each of the years that I have DWA data, I calculated summary statistics for each of the six traits. These statistics provided information concerning the distributions of the scores, the stability of each of the traits for all students and for students in the targeted groups across several years. Third, summary statistics from the DWA were compared to similar summary statistics from The National Assessment for Education Progress (NAEP) for Writing 2007. Fourth, I examined the reliability of the scores from the DWA. Fifth, I examined the validity of the DWA, taking a four-pronged approach. In this approach, I collected various kinds of empirical evidence and theoretical rationales to evaluate the proposed interpretations of the test scores in light of the purpose of the DWA. For the first prong of my examination of validity, I described the meaning of each trait and examined other writing assessments for evidence of these same traits. For the second prong, I examined the extent to which the six traits are used in instruction and assessment in schools across the United States. For the third prong, using exploratory and confirmatory factor analyses, I analyzed the internal structure of the DWA. Finally, for the fourth prong, I compared the results of my

confirmatory factor analyses across all 4 years of data and across ethnic and ELL groups. The degree to which evidence of validity based on one situation can generalize to other situations is reflective of the test's overall validity.

Published Information About the DWA

Pearson Publication Company was contracted by the Utah State Office of Education to score the DWA over the years 2004 to 2007. Pearson's published information focuses almost entirely on the scoring procedures of the DWA. They report interrater reliability as the percentage of exact and adjacent score agreement among raters. For the 2008 data, Pearson reports that exact and adjacent score point agreement was greater than 98% for each of the six traits. Because the same scoring procedures were followed for the 2004 to 2007 time periods, this value was likely consistent across those years. For holistic or analytic writing assessment, this value represents high interrater agreement.

Summary Statistics, Correlations, and Reliability

The mean scores for each of the six traits were consistently just above 3.0 for each year, and standard deviations were consistent. Correlations among the traits also were consistent and moderate to strong across the 4 years. Intraclass correlations among the six traits ranged from .72 to .80. Skewness and kurtosis were acceptable, with a slight negative skew across the 4 years (i.e., students' scores ended to cluster around 3.5 to 4.0). Cronbach Alpha was either .94 or .96 for the six traits for the 4 years, indicating strong interitem reliability.

In my calculations of the corrected item-total correlations and squared multiple correlations for each year, each trait was strongly correlated to the scale composed of the other traits. Across the 4 years and the six traits, organization had the highest correlation with the other items, accounting for about 83% in the other scores. Although writing instructors would not want to focus their attention exclusively on organization during instruction and ignore the other five traits; however, for quick scoring of students' writing samples, instructors could score the essays only on organization and have a fairly solid estimate of the other traits.

Comparison With NAEP

The goal of comparing results from the four administrations of the DWA to results from three administrations of the NAEP was to compile additional empirical evidence that results from the DWA converge with results from a large, nationally recognized writing assessment. If results from the two assessments diverge, questions would be raised about whether scores from the well-established NAEP are measuring aspects of writing that are purportedly being measured by the DWA, and therefore, the interpretation of the DWA scores would need to be questioned. For example, if the DWA scores showed no differences between males and females in writing skill, a fact that has been long established by the NAEP, as well as many other writing assessments, the scores from the DWA would be questionable.

Gains in Writing Performance

The NAEP has shown consistent gains in writing scores for eighth-grade writers across the United States from 1998 to 2002 to 2007. For the DWA, Utah's ninth-grade students showed gains only in 2007; however, the effect size was very small (i.e., .001). Unfortunately, this lack of convergence between the two assessments may indicate regional differences in writing skill development.

Writing Performance by Gender

The NAEP from all 3 years consistently showed that females outperformed males. The DWA from all 4 years showed that same effect, with females outperforming males, and this was true for all ethnic groups.

Writing Performance by Ethnicity

Both the NAEP and the DWA show complex patterns of writing scores of racial or ethnic groups across the years of administrations of the tests. For the national samples represented by the NAEP and the state samples represented by the DWA, white students tend to outperform other racial or ethnic groups. In Utah, however, Asian students score either slightly higher or on a par with White students. A consistent pattern shown by both assessments is that African American, Hispanic, and American Indian students tend to score lower than other racial/ethnic groups.

Writing Performance by Income

Statements concerning income for both the NAEP and the DWA are based on whether students are eligible for free or reduced-price lunch. For all national samples represented by the NAEP and all state sample represented by the DWA, students on free or reduced-price lunch scored lower than students who were not eligible for either. This finding is not restricted to writing performance, but generally is true of all areas of academics.

Writing Performance by ELL Status

Writing scores based on ELL status is not reported in the NAEP. However, the DWA provides these scores. A relatively straightforward hypothesis that could be made about ELL status and writing skill is that with lower English proficiency lower writing scores will result. With the exception of pre-emergent writers from 2004, this hypothesis was born out by the data across the 4 years. Greater proficiency in English is associated with greater proficiency in writing skill.

Validity

My examination of validity involved a four-pronged approach: I compared the meaning of each trait against writing constructs used by two other writing assessments, examined the extent to which the six traits are used; conducted exploratory and confirmatory factor analyses to analyze the internal structure of the DWA; and, compared the internal structure across all years of the DWA data and across ethnicity and ELL status.

Comparison of Six Traits With Other Writing Constructs

Essentially, the question I addressed by this analysis was whether writing skill is conceptualized similarly by the DWA and two other commonly used writing assessments, the SSQS and the College Board SAT Writing Exam. The analysis showed that three of the six traits, word choice, conventions, and organization, match closely with words, mechanics, and organization, respectively, from the SSQS, and with vocabulary, mechanics, and organization, respectively, from the SAT. The three assessments also matched on ideas/content/ideas; however, the SAT includes an additional element of organization and coherence for ideas, and therefore, overlaps to some degree with organization from the DWA and SSQS. The three assessments also matched on sentence fluency/style, sentence structure, although the SSQS includes a focus on being daring in the use of unique idea, which overlaps to some degree with voice in the DWA and point of view in the SAT. Finally, there was some divergence from voice as conceptualized by the six traits and writing constructs in the SSQS and SAT. Voice in the DWA focuses on whether the unique qualities of the author stand out in the writing. Purpose.audience.tone of the SSQS includes unique author qualities but also focuses on whether there is a clear purpose stated in the writing. The SAT also includes the unique author qualities but also includes whether the author has shown critical thinking. Although there are some divergences in the conceptualizations of writing as proposed by the three writing assessments, there are many more convergences. The definitive answer would be forthcoming from further research in which writing is assessed by all three assessment systems; however, at least conceptually, the definition of the construct of writing has more in common across the three writing assessments than what differs.

Acknowledged Use of Six Traits

The evidence gathered to answer whether the six traits are commonly used throughout the United States showed that they are used extensively in state assessment programs and in school district writing curricula. Moreover, they are also being used to some extent by teachers from South America, Africa, and the Far East. Clearly, conceptualizing writing as described by the six traits is widely acknowledged and used.

Internal Structure of the DWA

Exploratory and confirmatory factor analyses showed that for persuasive writing, the six traits reduce to three factors, which I have labeled patterned ideas, writer's perspective, and fluency components. In addition, adding a path from patterned ideas to voice added significantly to the model fit. Patterned ideas includes both ideas and organization. The ideas (i.e., meaning) contained within a persuasive essay, as with any essay, are the very heart of the essay, but how those ideas are organized within the essay will impact the overall meaning. In addition, with a persuasive essay, personal voice plays a significant role in the ideas that are generated and how they are organized. With the definition of persuasion as "An argument, appeal, etc., intended to induce belief or action; a means of persuading someone" (Oxford English Dictionary, http://dictionary.oed.com), the role of personal voice is very salient in both the ideas and the organization of those ideas. Moreover, recall that voice was measured consistently higher across all years and groups of the DWA data. Therefore, for persuasive essays, a definition for the factor of patterned ideas would be the personal ideas contained within

an essay and how those ideas are patterned or structured throughout the essay to persuade someone of a specific belief or action.

The second factor, writer's perspective, includes both voice and word choice. Voice, defined as the unique perspective of the writer, and word choice, defined as the use of language that is rich, colored, and moves the reader, are conceptually linked in a persuasive essay. The unique perspective (i.e., voice) of the writer in attempting to persuade others is expressed through his or her choice of words. Therefore, an appropriate definition for this factor would be the choice of words in an essay that reveal the writer's unique perspective on the focal topic of the essay.

Last, the third factor, fluency components, includes both conventions and sentence fluency. Conventions is defined as the appropriate placement of commas, proper use of grammar and usage, paragraphing, and punctuation; and sentence fluency is defined as the flow of language and sound of word patterns. Conceptually, the flow of language and word patterns would depend on such conventions as where commas are placed, the use of grammatical structures, word usage, and punctuation. Therefore, an appropriate definition for this factor would be the flow of language within an essay and the conventions that contribute to that flow.

This three-factor model of persuasive writing was validated across all 4 years of the DWA data and across ethnic groups and ELL statuses. Because of the high correlations among the six traits for all 4 years of the data and for ethnic and ELL groups, clearly there is a great deal of redundancy in what the six traits are measuring for persuasive essays. This redundancy can be substantially reduced by modeling persuasive writing with three traits: patterned ideas, writer's perspective, and fluency components.

Although there are still high correlations among these three traits, the three-trait model provides a more parsimonious rubric on which to measure writing quality for persuasive essays. This could aid considerably in the task of assessing persuasive essays, but perhaps more important, the three-trait model provides a more precise way of conceptualizing persuasive writing.

Using the six traits to assess all kinds of writing is something of a "one size fits all" approach to the assessment of writing. Different kinds of writing serve different purposes and involve different kinds of cognitive processing. There are some kinds of writing for which all of the six traits would not be appropriate. For example, for informative writing, voice would not be an important quality to be considered, and yet in persuasive writing it is very important. These differences in genre may require differences in the ways they are assessed. The six traits may be a good place to start, but analyses similar to the ones conducted here may provide more parsimonious and more precise descriptions of each genre, and more valid ways of assessing each.

Summary

Because reliability is a necessary but insufficient criterion for validity, I first examined the reliability of the scoring and then the scores of the DWA. The information provided by Pearson Publishing Company indicates that the scoring of the DWA is highly reliable, with a combined rate of exact and adjacent score point agreements greater than 98% for each of the six traits. The reliability of the scores as measured by Cronbach Alpha for all years and all ethnic and ELL groups was in excess of .90. Moreover, there

was strong consistency in mean scores, standard deviations, correlations, and intraclass correlations across all years of the DWA data.

Validity was addressed by gathering multiple sources of theoretical and empirical evidence to build a case that the test scores from the DWA allow appropriate inferences and decisions. The scores of the DWA across all 4 years show similar trends that have been identified by the NAEP, a large national assessment of writing as far as gender, ethnicity, and income are concerned. The scores follow intuitive patterns by ELL status, with greater English proficiency associated with greater writing skill. The conceptualizations of the six traits correspond strongly with the conceptualizations of writing that are used in two other widely recognized writing assessments. Further, the six traits used by the DWA are nationally and internationally recognized as valid measures of writing assessment. Last, the analyses of the internal structure of the DWA in the scoring of persuasive essays provide a structure that is consistently supported across all 4 years of the DWA data and for all ethnic and ELL groups. In sum, the empirical and theoretical evidence concerning the reliability and validity of the DWA that have been examined here support the use of the DWA test scores in making inferences and decisions about writing skill.

Research Questions 2, 3, and 4

2) What is the relation between ELL status and writing proficiency for ninth-grade students attending public schools in Utah during the years 2004, 2005, 2006, and 2007, and to what extent do student variables, gender, social economic status, and ethnicity independently and cumulatively explain the relation?

3) To what extent do the school variables, percent low-income students in a school, percent minority students in a school, size of the school, and mean ELL status at a school independently and cumulatively explain the relation between ELL status and writing proficiency?

4) To what extent do the district variables, percent low-income students in a district, percent minority students in a district, size of the district, mean ELL status in a district, and whether a district is urban or rural independently and cumulatively explain the relation between ELL status and writing proficiency?

These three research questions were answered by conducting a HLM analysis for each of the 4 years of DWA data using a three-level model, with student at level 1, school at level 2, and district at level 3. In addition to fixed effects for level 1 variables for each year, I found consistent contributions across the years from level 2 and level 3 variables.

Fixed Effects for Level 1 Variables

For all 4 years, there were fixed effects for level-one variables, gender, low-income, and ELL status; and for all 4 years except 2005, there was a fixed effect for the level one variable, White/non-White. The lack of a fixed effect for the variable White/non-White in 2005 was likely due to the absence of data for four urban and four rural schools for that year. These fixed effects were not unexpected because they directly reflect the descriptive statistics that were provided in response to research question one, and they reflect trends that have been found in national data sets (e.g., NAEP). For the 4 years, females outperformed males, students who did not receive financial support for lunch outperformed students who did, and students at higher ELL statuses outperformed

students at lower ELL statuses. And, with the exception of 2005, White students outperformed non-White students.

Contributions From ELL Status at School and District Levels

The contributions from ELL status were seen at both school and district levels across 3 of the 4 years, once again, excluding 2005 with the loss of a large proportion of ELL students from the four missing urban districts. For the 2004, 2006, and 2007 data, the level two variable, number of ninth-grade students in a school, moderated the effect of ELL status on writing proficiency. With an increase in the number of ninth-grade students in a school, the stronger the relationship between ELL status and writing proficiency became. That is, differences due to ELL status are stronger at larger schools than at smaller schools. This may be due to the greater diversity in ELL students at larger schools. With larger populations, there is a greater chance that differences in ELL status will be larger. The greater diversity in ELL statuses will strengthen the relationship between ELL status and writing proficiency, such that differences in writing proficiency become greater with greater diversity. Although larger schools tend to have more resources for ELL students, such as more ESL programs and ESL certified teachers, there still exist stronger differences between the ELL status and writing proficiency than in smaller schools. This may be influenced strongly by students at the lower levels of English proficiency who show the lowest levels of writing proficiency.

School effects of ELL status on writing proficiency also were evident in the 2007 data. The mean ELL status for the ninth-grade population at a school moderated the effect of ELL status at level one. That is, with greater English proficiency of ELL

students at a school, the relation between ELL status and writing proficiency lessens. This is a reasonable finding in that one would expect to see fewer differences in writing proficiency due to ELL status in schools in which the majority of ELL students are more proficient in English.

Finally, district effects of ELL status on writing proficiency were evident in the 2006 data. In districts with higher proportions of students on free or reduced lunch, differences in writing proficiency lessened among the ELL statuses. That is, the writing scores look more alike across differences in English language proficiency. At first glance, this may seem to be a positive finding in that districts with higher proportions of low-income students appear to have fewer differences in writing ability due to ELL status. However, even though the differences due to ELL status are lessened, because income plays a significant role in writing performance, the lessened differences are at a lower level of performance.

In sum, as expected, greater proficiency in English is associated with greater writing proficiency; however, this relation is not completely straightforward in that it changes due to differences in schools and districts. The effects of ELL status on writing proficiency are stronger in larger schools (i.e., schools with larger ninth-grade populations), and are weaker in schools in which the mean ELL status is higher and in districts in which there are higher proportions of students receiving assistance for lunch programs.

Contributions From Number of Students at School Level

For all 4 years, the number of ninth-grade students at a school independently contributed to writing proficiency. Schools with larger ninth-grade populations tended to score higher than schools with smaller ninth-grade populations. Number of ninth-grade students in a school also had a moderating effect on gender in the 2006 data. That is, the effects of gender were lessened in schools with larger ninth-grade populations. These findings are in need of further investigation. Schools with larger ninth-grade populations are likely larger schools overall. Larger schools may have greater resources available to help students academically, they may also have a larger offering of courses in which students could focus on writing. Although the effect of gender was lessened for the one year, females still outperformed males. Why for only 1 year this effect was lessened needs further examination.

Contributions From Proportion of Students on Free or Reduced Lunch Programs at School and District Levels

For 3 of the 4 years of data (2004, 2005, and 2007), the proportion of students at a school who received financial assistance for lunch programs contributed independently to writing proficiency. Schools with larger proportions of students receiving financial assistance tended to score lower than schools with smaller proportions of students receiving financial assistance. This finding was not unexpected given the overwhelming evidence that income plays a significant role in overall school performance. Low-income has adverse outcomes at the individual student level and at the school level.

Although the 2006 data did not show an effect of low income at the school level, low income showed an effect at the district level. Districts with higher proportions of students on free or reduced lunch showed less difference in writing proficiency among the ELL statuses. On one hand, this could be interpreted in a positive way in that ELL status seems to play less a role in districts with higher proportions of low-income students than in districts with lower proportions of low-income students. On the other hand, this also could be interpreted that ELL students from districts with higher proportions of low-income students score more alike, with the lower scoring students scoring higher, but with the higher scoring students scoring lower.

Proportion of low-income students at a school also had moderating effects in 2007 on the level one variable of gender, and moderating effects in 2005 on gender and low-income. For the gender effects in 2005 and 2007, as the proportion of students who received financial assistance for lunch programs at a school increased so did the relation between gender and writing proficiency. For 2005, as the proportion of students at a school who received free or reduced lunch increased, the effect of income on students' writing proficiency increased.

The latter of these moderation effects is fairly straightforward, that is, low-income students in schools with higher proportions of low-income students have lower writing proficiency than low-income students in schools with lower proportions of low-income students. Thus, although this effect appears only in the 2005 data, the adverse effects of low-income on writing proficiency for a student appear to be exacerbated in schools with higher proportions of low-income students.

The moderating effects of low-income on writing proficiency for males and females in years 2005 and 2007 are a bit more difficult to interpret. As the proportion of low-income students at a school increases, the differences between males and females increases. One interpretation of these findings is that schools with higher proportions of low-income students may have greater adverse affects on males than females. Conversely, females may be better able to thrive than males in low-income schools. Examining the differential effects on writing, and perhaps learning in general, caused by poverty is beyond the scope of the present study, but this does identify an important area for future research.

Contributions From Size of School District

For the 2006 and 2007 data, the size of the school district (measured by number of schools serving ninth-grade students) moderated the effects of ethnicity. The level 1 effect of ethnicity on writing proficiency indicated that White students outperformed non-White students; however, in larger school districts, differences between Whites and non-Whites increased. An interpretation of this finding is that there is likely greater ethnic diversity in larger school districts, that is, a greater percentage of non-White students. Similar to the findings reported in the NAEP Report on Writing (2007) in which it was stated that deficiencies in writing skills are at their greatest when ethnicity, social economic status, and limited English language proficiency are considered, in school districts with larger percentages of non-White students, the disparity between Whites and non-Whites widens.

This increased disparity between Whites and non-Whites also was evident at the school level in the 2006 data. The proportion of non-White students in a school contributed independently to writing proficiency: With greater ethnic diversity in a school, greater differences in writing proficiency appeared between Whites and non-Whites.

Research Question 5

5) To what extent do each of the six traits vary with ELL status and ethnicity, and are there interactions between ELL status and ethnicity on any of the six traits?

The analyses of ethnic group and ELL status showed discernible patterns in writing skill. In general, for the lower three ELL statuses, there was consistent growth in writing skill from pre-emergent to the intermediate statuses: Increased English language proficiency was matched with increased writing skill. The one exception to this was with White ELL students, who showed no growth between these lower ELL statuses. Exactly who these White ELL students were and why their writing skill did not grow with increases in English language skill are questions for future research.

Writing skill began to vary at the higher ELL statuses for the ethnic groups. Between exited and advanced statuses, Asian, White, Hispanic, and Pacific Islander students showed significant growth in writing skill for all six traits. In general, exited students from these four ethnic groups were writing on a par with non-ELL students. For Native Americans and African Americans there was little or no growth in writing skill between these two higher ELL statuses, and for the African Americans, there were no differences in writing skill between the three higher ELL statuses. For African

Americans, after the growth in English proficiency and writing skill from pre-emergent to intermediate-advanced, their writing skill plateaued.

When examining the six ethnic groups without considering ELL status, there exists a writing achievement gap among them. Asians outperformed Whites, who outperformed Pacific Islanders, who outperformed African Americans, who outperformed Hispanics, who outperformed American Indians. Within each ethnic group, there were differences in performance on each of the six traits. All groups did best with voice, which may have been a consequence of writing a persuasive essay. However, Asian and Pacific Islander students scored lowest on word choice; African Americans scored lowest on conventions; Whites scored lowest on sentence fluency, conventions and word choice; Hispanics scored lowest on sentence fluency, conventions, word choice, and organization; and finally, American Indians scored lowest on sentence fluency, conventions, organization, and ideas. In sum, most of the groups show a slightly different pattern of traits that need instructional attention, with some groups needing more attention than others.

Whether these differences exist because of cultural differences remains to be investigated. All we know is that when measured by the same yardstick (i.e., the six traits), the differences among ethnic groups are evident. This yardstick was shown to produce reliable and valid measures for the population of ninth-grade students in Utah. The degree to which a student's or a specific group's score varies from a given measure on that yardstick indicates the degree to which that student or group of students has or has not acquired a specific height. Whether people choose to accept that yardstick as a standard measure is a choice to be made with deliberation. In the 1990s, there were many

arguments made about recognizing standard English across all ethnic groups. A point well made at the time was that standard English is the language of power in the United States, and that other forms of English do not allow their speakers full access to that power. Recognizing a specific standard of writing as the written language of power in the United States may need similar arguments to be made.

Limitations of the Study

Important limitations of this study begin with the realization that only one genre of writing was assessed - persuasive essays. Future studies could analyze other genres of writing such as descriptive, expository, or narrative. In this way, even though the traits assessing the writing would remain the same, assessing different forms of writing could further validate the results rendered from this study.

Another limitation was the set of data used. While the sample size was extensive, it only measured ninth-grade students in the State of Utah. The same assessment used to analyze the writing – the DWA – is used in other states as a writing assessment tool as well. A comparison of the results from other states could give an indication as to how well the students in Utah are doing with writing at a national level.

Because the DWA is an analytic writing assessment that uses six traits, other forms of writing assessments such as discrete point assessment or holistic assessment could be used and correlated with scores from the six traits. High correlations would provide further validation of the use of the six traits as a measure of writing proficiency. Other assessments of writing would analyze the writing of ELL and ethnicity students differently. A final limitation would be that ELL status could vary throughout other

states. The way students' ethnicity is identified for assessment purposes varies in other states.

Implications and Future Research

Implications from this study are primarily in the area of instruction and assessment. From this study, I have shown that differences exist in writing skill among different ethnic groups and ESL statuses. If use of the DWA, which makes use of the six traits, is to remain the standard of measurement for writing, the challenge remains of how best to serve non-White students and students who are at lower levels of English language proficiency. Although Asians are on a par or slightly above White students in writing proficiency, there are large differences between Whites and Pacific Islanders, African Americans, Hispanic, with American Indian students showing the least writing proficiency. Although these disparities may have a cultural origin, much greater effort needs to be put forth to get these students on a par with White and Asian students. Moreover, when ELL status is considered, these efforts become more complex because the strengths and deficiencies in writing skills vary among ethnicity and ELL status. These efforts may involve changes in writing curriculum, in amount of instructional time, in better trained teachers, or perhaps even a change in assessment methods. These are decisions that the educational leadership in Utah must address if *all* students are to be given an opportunity to share equally in the benefits from mastering a standard written language in the United States.

Another implication is determining whether there are better ways to assess student writing. Although the six traits provide a standard of writing that is well-recognized, the

results of this study show that with the high correlations among the traits there is a great deal of redundancy in the scores. What sense does it make to measure children's height using a yardstick only to measure them again using a meter stick? By only using organization as a trait of assessment, the results showed that 83% variability in the other traits could be explained. This does not mean that only organization should be taught to children. However, as a quick assessment of persuasive writing, what sense does it make to measure children's writing skill using all six traits when just one trait will provide nearly the same information?

Relatedly, exploratory and confirmatory factor analyses showed that for persuasive writing the six traits could be combined into three, which I described and labeled as patterned ideas, writer's perspective, and fluency components. I offer that these three may be a better way of conceptualizing persuasive writing, and further, that they would provide a more parsimonious way of assessing persuasive writing.

Larger schools appeared to have substantial impact on students' writing proficiency. Larger schools—and for 1 year larger districts—tend to have higher writing scores than smaller schools, display greater disparity in writing skill between Whites and non-Whites, and have bigger differences between the ELL statuses. Although some of these findings simply could be the result of greater diversity in students in larger schools, these findings do warrant future examinations into the differences between larger and smaller schools. Larger schools tend to have more resources for ELL students, such as more ESL programs and ESL certified teachers, and they tend to have a greater array of course offerings than smaller schools. On one hand, there definitely appear to be some positive associations between these resources and writing proficiency, but on the other

hand, despite these advantages of larger schools, significant differences due to ethnicity and ELL status exist. Examining the causes of these unfortunate differences could reveal some important implications for the kinds of writing curricula that are provided at different size schools and whether differential instruction of writing should be considered.

Last, as far as actual writing instruction is concerned, several implications have been suggested by this research involving differential writing instruction. Specifically, writing instruction for Hispanics should focus more strongly on teaching conventions; Whites and Pacific Islanders would benefit by increasing the emphasis of writing instruction in the area of word choice; both American Indian and African American students would benefit from increased instruction in the area of conventions. The teaching of organization would also be an area in which American Indians need additional instruction. Asians, as the top overall performing students, could excel even more with added emphasis on word choice.

Conclusion

Writing has remained the "Neglected R" in the traditional triad of what students need to learn (IES, 2003). Moreover, the deficiencies in writing skills are at their greatest when ethnicity, social economic status, and limited English language proficiency are considered. The inability to write effectively is likely to have an impact on the future academic prospects and employment of students. Students who are unable to write and communicate effectively stand the chance of not being hired, not being admitted to institutions of higher education, and of not being able to participate equally in the power

bases of today's society. Perhaps even more disheartening, by not assisting students to become the best writers they can be, we are preventing them from realizing higher levels of intellectual achievement.

As educators, we face the challenge of preparing students to use writing as a way to communicate, to learn, persuade others, and as a means for personal self-expression. Although we still have a long way to go to meet this challenge, more and more educators and political leaders are beginning to realize that writing is central to success in and out of school (National Writing Project & Nagin, 2006). The results of the present study have suggested ways to rethink how writing is conceptualized and assessed, how past instructional practices have possibly resulted in disparities among ethnic groups, how writing skills vary with student, school, and district characteristics, and how differential writing instruction may benefit students of different ethnic groups and ELL statuses. Because Utah has been identified by the Bureau of Economic and Business Research (2003) as one of the fastest changing states in the nation in terms of diversity, with a projected increase of 278% minority growth by 2025, we have the need for educators to understand and apply current research, especially in the area of writing. If teachers are to be effective in helping all students increase their writing proficiency—especially to under-represented groups—they need to understand how to differentiate their writing instruction. A "one size fits all" approach to writing instruction will not benefit all students in Utah or throughout the nation. The results of this study show that writing does differ along ELL and ethnicity lines; therefore, writing instruction must differ as well.

APPENDIX A

THE SIX TRAITS AND SCORING CRITERIA

Ideas

The heart of the message, the content of the piece, the main theme, with details that enrich and develop that theme.

5 points – This paper is clear and focused. It holds the reader's attention. Relevant anecdotes and details enrich the central theme.

A. The topic is narrow and manageable.

B. Relevant, telling, quality details give the reader important information that goes beyond the obvious or predictable.

C. Reasonably accurate details are present to support the main ideas.

D. The writer seems to be writing from knowledge or experience; the ideas are fresh and original.

E. The reader’s questions are anticipated and answered.

F. Insight—an understanding of life and a knack for picking out what is significant—is an indicator of high-level performance, though not required.

3 points – The writer is beginning to define the topic, even though development is still basic or general.

A. The topic is fairly broad; however, you can see where the writer is headed.

B. Support is attempted but doesn't go far enough yet in fleshing out the key issues or story line.

C. Ideas are reasonably clear, though they may not be detailed, personalized, accurate, or expanded enough to show in-depth understanding or a strong sense of purpose.

D. The writer seems to be drawing on knowledge or experience but has difficulty going from general observations to specifics.

E. The reader is left with questions. More information is needed to "fill in the blanks."

F. The writer generally stays on the topic but does not develop a clear theme. The writer has not yet focused the topic past the obvious.

1 point – As of yet, the paper has no clear sense of purpose or central theme. To extract meaning from the text, the reader must make inferences based on sketchy or missing details. The writing reflects more than one of these problems:

A. The writer is still in search of a topic, brainstorming, or has not yet decided what the main idea of the piece will be.

B. Information is limited or unclear or the length is not adequate for development.

C. The idea is a simple restatement of the topic or an answer to the question with little or no attention to detail.

D. The writer has not begun to define the topic in a meaningful, personal way.

E. Everything seems as important as everything else; the reader has a hard time sifting out what is important.

F. The text may be repetitive or may read like a collection of disconnected, random thoughts with no discernable point.

Organization

The internal structure, the thread of central meaning, the logical and sometimes intriguing pattern of the ideas.

5 points – The organization enhances and showcases the central idea or theme. The order, structure, or presentation of information is compelling and moves the reader through the text.

A. An inviting introduction draws the reader in; a satisfying conclusion leaves the reader with a sense of closure and resolution.

B. Thoughtful transitions clearly show how ideas connect.

C. Details seem to fit where they're placed; sequencing is logical and effective.

D. Pacing is well-controlled; the writer knows when to slow down and elaborate, and when to pick up the pace and move on.

E. The title, if desired, is original and captures the central theme of the piece.

F. The choice of structure matches the purpose and audience, with effective paragraph breaks.

3 points – The organizational structure is strong enough to move the reader through the text without too much confusion.

A. The paper has a recognizable introduction and conclusion. The introduction may not create a strong sense of anticipation; the conclusion may not tie up all loose ends.

B. Transitions sometimes work; at other times, connections between ideas are unclear.

C. Sequencing shows some logic, but not under control enough that it consistently supports the development of ideas. The structure may be predictable and taking attention away from the content.

D. Pacing is fairly well-controlled, though the writer sometimes lunges ahead too quickly or spends too much time on details that do not matter.

E. A title (if desired) is present, although it may be uninspired or an obvious restatement of the prompt or topic.

F. The organization sometimes supports the main point or story line, with an attempt at paragraphing.

1 point – The writing lacks a clear sense of direction. Ideas, details, or events seem strung together in a loose or random fashion; there is no identifiable internal structure. The writing reflects more than one of these problems:

A. There is no real lead to set up what follows, no real conclusion to wrap things up.

B. Connections between ideas are confusing or absent.

C. Sequencing is random and needs lots of work.

D. Pacing feels awkward; the writer slows to a crawl when the reader wants to move on, and vice versa.

E. No title is present (if requested) or, if present, does not reflect the content.

F. Problems with organization make it hard for the reader to understand the main point or story line, with little or no attempt at paragraph breaks.

Voice

The unique perspective of the writer coming through honesty, conviction, and integrity and believability.

5 points –The writer speaks directly to the reader in a way that is individual, compelling, and engaging. The writer crafts the writing with an awareness and respect for the audience and the purpose for writing.

A. The writer connects strongly with the audience through the intriguing focus of the topic, selection of relevant details, and the use of natural, engaging language.

B. The purpose of the writing is accurately reflected in the writer's choice of individual and compelling content, and the arrangement of ideas.

C. The writer takes a risk by the inclusion of personal details that reveal the person behind the words.

D. Expository or persuasive writing reflects a strong commitment to the topic by the careful selection of ideas that show why the reader needs to know this.

E. Narrative writing is personal and engaging, and makes the reader think about the author's ideas or point of view.

3 points – The writer seems sincere but not fully engaged or involved. The writing has discernable purpose but is not compelling.

A. The writing attempts to connect with the audience in an earnest, pleasing, but impersonal manner.

B. The writer seems aware of a purpose and attempts to select content and structures that reflect it.

C. The writer occasionally reveals personal details but primarily avoids risk.

D. Expository or persuasive writing lacks consistent engagement with the topic and fails to use ideas to build credibility.

E. Narrative writing is sincere but does not reflect a unique or individual perspective on the topic.

1 point – The writer seems indifferent to the topic and the content. The writing lacks purpose and audience engagement.

A. The writer's ideas and language fail to connect with the audience.

B. The writer has no clear purpose, and the chosen style does not match the content or ideas.

C. The writing is risk-free and reveals nothing about the author.

D. Expository or persuasive writing is lifeless and mechanical, or lacks accurate information.

E. Narrative: The development of the topic is so limited that no point of view is discernable.

Word Choice

The use of rich, colored, precise language that moves and enlightens the reader.

5 points –Words convey the intended message in a precise, interesting, and natural way. The words are powerful and engaging.

A. Words are specific and accurate. It is easy to understand just what the writer means.

B. Striking words and phrases often catch the reader's eye and linger in the reader's mind.

C. Language and phrasing are natural, effective, and appropriate for the audience.

D. Lively verbs add energy, while specific nouns and modifiers add depth.

E. Choices in language enhance the meaning and clarify understanding.

F. Precision is obvious. The writer has taken care to put just the right word or phrase in just the right spot.

3 points – The language is functional, even if it lacks much energy. It is easy to figure out the writer's meaning on a general level.

A. Words are adequate and correct in a general sense, and they support the meaning by not getting in the way.

B. Familiar words and phrases communicate but rarely capture the reader's imagination.

C. Attempts at colorful language show a willingness to stretch and grow but sometimes reach beyond the audience (thesaurus overload!).

D. Despite a few successes, the writing is marked by passive verbs, everyday nouns, and mundane modifiers.

E. The words and phrases are functional with only one or two fine moments.

F. The words may be refined in a couple of places, but the language looks more like the first thing that popped into the writer’s mind.

1 point – The writer demonstrates a limited vocabulary or has not searched for words to convey specific meaning.

A. Words are so nonspecific and distracting that only a very limited meaning comes through.

B. Problems with language leave the reader wondering. Many of the words just don't work in this piece.

C. Audience has not been considered. Language is used incorrectly, making the message secondary to the misfires with the words.

D. Limited vocabulary and/or misused parts of speech seriously impair understanding.

E. Words and phrases are so unimaginative and lifeless that they detract from the meaning.

F. Jargon or clichés distract or mislead. Redundancy may distract the reader.

Sentence Fluency

The rhythm and flow of the language, the sound of word patterns, the way in which the writing plays to the ear, not just to the eye.

5 points – The writing has an easy flow, rhythm, and cadence. Sentences are well-built, with strong and varied structure that invites expressive oral reading.

A. Sentences are constructed in a way that underscores and enhances the meaning.

B. Sentences vary in length as well as structure. Fragments, if used, add style. Dialogue, if present, sounds natural.

C. Purposeful and varied sentence beginnings add variety and energy.

D. The use of creative and appropriate connectives between sentences and thoughts shows how each relates to, and builds upon, the one before it.

E. The writing has cadence; the writer has thought about the sound of the words as well as the meaning.

F. The first time the work is read aloud is a breeze.

3 points – The text hums along with a steady beat but tends to be more pleasant or businesslike than musical, more mechanical than fluid.

A. Although sentences may not seem artfully crafted or musical, they get the job done in a routine fashion.

B. Sentences are usually constructed correctly; they hang together; they are sound.

C. Sentence beginnings are not all alike; some variety is attempted.

D. The reader sometimes has to hunt for clues (e.g., connecting words and phrases such as: however, therefore, naturally, after a while, on the other hand, to be specific, for example, next, first of all, later, but as it turned out, although, etc.) that show how sentences interrelate.

E. Parts of the text invite expressive oral reading; others may be stiff, awkward, choppy, or gangly.

1 point – The reader has to practice quite a bit in order to give this paper a fair interpretive reading. The writing reflects more than one of the following problems:

A. Sentences are choppy, incomplete, rambling or awkward; they need work. Phrasing does not sound natural. The patterns may create a sing-song rhythm, or a chop-chop cadence that lulls the reader to sleep.

B. There is little to no "sentence sense" present. Even if this piece was flawlessly edited, the sentences would not hang together.

C. Many sentences begin the same way—and may follow the same patterns (e.g., *subject-verb-object*) in a monotonous pattern.

D. Endless connectives (*and, and so, but then, because, and then,* etc.) or a complete lack of connectives create a massive jumble of language.

E. The text does not invite expressive oral reading.

Conventions

The mechanical correctness of the piece; spelling, grammar and usage, paragraphing, use of capitals and punctuation.

5 points – The writer demonstrates a good grasp of standard writing conventions (e.g., spelling, punctuation, capitalization, grammar, usage, paragraphing) and uses conventions effectively to enhance readability. Errors tend to be so few that just minor touch-ups would get this piece ready to publish.

A. Spelling is generally correct, even on more difficult words.

B. The punctuation is accurate, even creative, and guides the reader through the text.

C. A thorough understanding and consistent application of capitalization skills is present.

D. Grammar and usage are correct and contribute to clarity and style.

E. Paragraphing tends to be sound and reinforces the organizational structure.

F. The writer may manipulate conventions for stylistic effect—and it works! The piece is very close to being ready to publish.

3 points – The writer shows reasonable control over a limited range of standard writing conventions. Conventions are sometimes handled well and enhance readability; at other times, errors are distracting and impair readability.

A. Spelling is usually correct or reasonably phonetic on common words, but more difficult words are problematic.

B. End punctuation is usually correct; internal punctuation (commas, apostrophes, semicolons, dashes, colons, parentheses) is sometimes missing/wrong.

C. Most words are capitalized correctly; control over more sophisticated capitalization skills may be spotty.

D. Problems with grammar or usage are not serious enough to distort meaning but may not be correct or accurately applied all of the time.

E. Paragraphing is attempted but may run together or begin in the wrong places.

F. Moderate editing (a little of this, a little of that) would be required to polish the text for publication.

1 point – Errors in spelling, punctuation, capitalization, usage, and grammar and/or paragraphing repeatedly distract the reader and make the text difficult to read. The writing reflects more than one of these problems:

A. Spelling errors are frequent, even on common words.

B. Punctuation (including terminal punctuation) is often missing or incorrect.

C. Capitalization is random and only the easiest rules show awareness of correct use.

D. Errors in grammar or usage are very noticeable, frequent, and affect meaning.

E. Paragraphing is missing, irregular, or so frequent (every sentence) that it has no relationship to the organizational structure of the text.

F. The reader must read once to decode, then again for meaning. Extensive editing (virtually every line) would be required to polish the text for publication.

APPENDIX B

2004 DESCRIPTIVE DATA FOR TOTAL SCORE ON THE DWA

Asian (*n* = 584)										
	Female (*n* = 283)					Male (*n* = 301)				
	n	Mean (*SD*)	Med	Min	Max	*n*	Mean (*SD*)	Med	Min	Max
Pre	1	13.00	13.00	13	13	3	15.67 (1.53)	16.00	14	17
Beg	19	16.53 (3.78)	18.00	9	24	22	16.54 (3.75)	16.50	9	24
Inter	7	19.86 (2.19)	20.00	18	24	11	19.00 (1.90)	19.00	16	22
Adv	27	22.15 (4.09)	24.00	12	30	46	20.78 (3.15)	20.50	15	30
Exit	60	23.47 (3.17)	24.00	15	30	62	22.00 (3.52)	22.00	16	30
Non	169	23.30 (3.78)	24.00	12	30	157	21.94 (3.62)	22.00	12	30

White (*n* = 28946)										
	Female (*n* = 14200)					Male (*n* = 14746)				
	n	Mean (*SD*)	Med	Min	Max	*n*	Mean (*SD*)	Med	Min	Max
Pre	4	18.25 (2.06)	18.00	16	21	3	18.00 (2.00)	18.00	16	20
Beg	33	19.03 (4.01)	18.00	11	26	25	15.92 (3.55)	18.00	7	22
Inter	6	16.50 (3.27)	17.00	11	20	16	18.94 (3.40)	18.00	12	24
Adv	48	20.85 (3.15)	21.00	14	30	32	19.03 (3.19)	19.00	10	25
Exit	51	22.71 (3.28)	23.00	15	30	45	23.31 (3.87)	22.00	16	30
Non	14058	22.38 (3.40)	23.00	6	30	14625	20.95 (3.72)	21.00	6	30

Hispanic (*n* = 3224)										
	Female (*n* = 1563)					Male (n = 1661)				
	n	Mean (*SD*)	Med	Min	Max	*n*	Mean (*SD*)	Med	Min	Max
Pre	46	18.89 (3.91)	19.00	7	28	52	16.96 (3.96)	18.00	7	25
Beg	237	17.41 3.21)	18.00	9	25	306	16.03 (3.60)	16.00	6	28
Inter	125	18.46 (2.89)	18.00	9	27	145	17.24 (3.42)	18.00	6	30
Adv	264	19.99 (3.05)	20.00	6	29	267	18.88 (3.23)	18.00	10	29
Exit	167	21.13 (3.42)	21.00	12	30	176	20.30 (3.56)	20.00	8	30
Non	724	20.42 (3.55)	20.00	9	30	715	19.25 (3.90)	19.00	6	30

Indian (*n* = 584)										
	Female (*n* = 291)					Male (*n* = 293)				
	n	Mean (*SD*)	Med	Min	Max	*n*	Mean (*SD*)	Med	Min	Max
Pre	1	24.00	24.00	24	24	0	0	0	0	0
Beg	46	16.41 (3.15)	17.50	10	24	60	14.35 (4.08)	14.00	6	24
Inter	21	18.43 (2.64)	18.00	13	24	31	15.97 (3.01)	16.00	10	24
Adv	35	19.57 (3.12)	19.00	12	24	34	18.44 (3.97)	18.00	10	27
Exit	19	20.10 (2.98)	21.00	12	24	15	18.80 (3.28)	18.80	13	25
Non	169	19.99 (3.40)	20.00	11	30	153	18.91 (4.08)	18.00	6	30

Pacific Islander (*n* = 440)										
	Female (*n* = 222)					Male (*n* = 218)				
	n	Mean (*SD*)	Med	Min	Max	*n*	Mean (*SD*)	Med	Min	Max
Pre	0	0	0	0	0	2	7.50 (2.12)	7.50	6	9
Beg	17	18.00 (2.48)	18.00	12	24	22	15.95 (3.09)	15.50	12	23
Inter	10	19.50 (1.43)	19.50	17	22	8	18.88 (2.30)	18.50	16	24
Adv	30	20.53 (3.35)	21.50	12	24	34	19.03 (3.47)	18.00	13	30
Exit	36	21.58 (3.75)	20.50	15	30	37	21.14 (2.87)	21.00	13	26
Non	129	22.14 (3.48)	22.00	11	30	115	20.19 (3.40)	20.00	10	28

African American (*n* = 330)										
	Female (*n* = 143)					Male (*n* = 187)				
	n	Mean (*SD*)	Med	Min	Max	*n*	Mean (*SD*)	Med	Min	Max
Pre	1	18.00	18.00	18	18	2	16.50 (.71)	16.50	16	17
Beg	5	17.40 (2.97)	17.00	14	22	9	15.67 (2.83)	16.00	12	20
Inter	3	20.33 (2.52)	20.00	18	23	2	15.50 (4.95)	15.50	12	19
Adv	3	20.00 (7.21)	18.00	14	28	8	20.12 (4.36)	21.00	13	27
Exit	3	22.67 (4.16)	24.00	18	26	3	21.33 (2.08)	22.00	19	23
Non	128	20.84 (3.47)	21.00	13	30	163	19.15 (3.76)	18.00	7	27

APPENDIX C

2005 DESCRIPTIVE DATA FOR TOTAL SCORE ON THE DWA

Asian (*n* =444)										
	Female (*n* = 212)					Male (*n* = 232)				
	n	Mean (*SD*)	Med	Min	Max	*n*	Mean (*SD*)	Med	Min	Max
Pre	1	18.00	18.00	18	18	5	16.60 (5.37)	18.00	10	22
Beg	13	18.54 (3.20)	18.00	13	25	15	17.73 (3.10)	18.00	11	24
Inter	7	20.71 (3.82)	19.00	17	26	13	19.69 (5.15)	19.00	12	29
Adv	25	23.20 (2.50)	24.00	18	28	30	21.00 (2.91)	20.50	17	27
Exit	66	24.05 (3.03)	24.00	16	30	63	22.44 (3.79)	24.00	12	30
Non	100	22.92 (3.50)	24.00	14	30	106	21.03 (4.09)	21.00	9	30

White (*n* =23089)										
	Female (*n* = 11198)					Male (*n* =11891)				
	n	Mean (*SD*)	Med	Min	Max	*n*	Mean (*SD*)	Med	Min	Max
Pre	2	24.00 (4.24)	24.00	21	27	1	16.00	16.00	16	16
Beg	9	17.22 (3.23)	19.00	13	20	11	16.36 (3.04)	16.00	12	23
Inter	6	17.50 (2.26)	17.50	14	21	11	18.82 (5.0)	20.00	9	24
Adv	27	22.19 (3.54)	24.00	17	29	29	19.72 (3.01)	19.00	12	24
Exit	60	23.20 (3.07)	24.00	15	30	57	21.81 (3.55)	22.00	13	28
Non	11094	22.30 (3.44)	23.00	7	30	11782	20.78 (3.82)	21.00	6	30

Hispanic (*n* = 2206)										
	Female (*n* = 1045)					Male (*n* =1161)				
	n	Mean (*SD*)	Med	Min	Max	*n*	Mean (*SD*)	Med	Min	Max
Pre	20	12.20 (5.37)	12.00	6	24	20	12.75 (3.29)	12.00	6	18
Beg	112	16.17 (3.92)	16.00	6	25	136	15.32 (3.65)	15.50	6	24
Inter	78	18.31 (3.30)	18.00	11	25	104	17.36 (3.87)	18.00	9	27
Adv	193	20.75 (3.28)	21.00	11	30	225	19.19 (3.38)	19.00	8	28
Exit	179	21.51 (3.32)	22.00	11	29	202	20.82 (3.69)	21.00	8	30
Non	463	21.29 (3.72)	22.00	7	30	474	19.37 (3.94)	19.00	6	30

Indian (*n* =453)										
	Female (*n* =246)					Male (*n* =207)				
	n	Mean (*SD*)	Med	Min	Max	*n*	Mean (*SD*)	Med	Min	Max
Pre										
Beg	39	16.46 (3.56)	17.00	8	24	55	16.00 (3.78)	17.00	6	24
Inter	10	18.80 (2.90)	19.00	14	24	19	16.74 (3.19)	17.00	9	24
Adv	30	19.80 (4.26)	19.50	8	27	19	18.89 (4.19)	18.00	8	25
Exit	43	20.60 (3.04)	20.00	16	30	18	20.22 (2.51)	19.50	17	24
Non	124	20.56 (3.61)	20.00	12	30	126	18.20 (3.93)	18.00	6	29

Pacific Islander (*n* = 538)										
	Female (*n* =182)					Male (*n* = 356)				
	n	Mean (*SD*)	Med	Min	Max	*n*	Mean (*SD*)	Med	Min	Max
Pre	1	22	22	22	22	1	12	12	12	12
Beg	8	18.50 (2.51)	18.00	16	24	18	14.44 (4.08)	13.50	6	24
Inter	4	22.25 (2.87)	23.50	18	24	4	18.50 (1.29)	18.50	17	20
Adv	18	22.28 (2.70)	24.00	18	26	17	20.65 (3.60)	19.00	15	27
Exit	43	22.70 (3.69)	24.00	7	29	35	21.63 (3.63)	23.00	12	28
Non	108	22.57 (3.20)	24.00	15	30	99	20.83 (3.62)	20.00	12	30

African American (*n* = 233)										
	Female (*n* = 112)					Male (*n* = 121)				
	n	Mean (*SD*)	Med	Min	Max	*n*	Mean (*SD*)	Med	Min	Max
Pre	0	0	0	0	0	0	0	0	0	0
Beg	1	16.00	16.00	16	16	13	15.23 (3.63)	17.00	7	19
Inter	0	0	0	0	0	2	14.00 (1.41)	14.00	13	15
Adv	0	0	0	0	0	3	18.67 (5.51)	16.00	15	25
Exit	1	25.00	25.00	25	25	3	18.67 (5.03)	18.00	14	24
Non	110	20.65 (3.83)	21.00	7	29	100	18.50 (4.22)	18.00	9	28

APPENDIX D

2006 DESCRIPTIVE DATA FOR TOTAL SCORE ON THE DWA

Asian (*n* =564)										
	Female (*n* =273)					Male (*n* = 291)				
	n	Mean (*SD*)	Med	Min	Max	*n*	Mean (*SD*)	Med	Min	Max
Pre	2	15.00 (9.90)	15.00	8	22	2	16.50 (2.12)	16.50	15	18
Beg	11	18.27 (2.69)	18.00	14	24	21	17.05 (3.58)	18.00	9	24
Inter	5	19.40 (3.13)	18.00	16	24	15	19.67 (2.66)	19	14	24
Adv	29	22.90 (3.26)	24.00	14	29	47	20.51 (4.0)	21.00	12	28
Exit	67	22.79 (2.92)	24.00	18	30	49	22.18 (3.29)	24.00	16	30
Non	159	23.37 (3.37)	24.00	15	30	157	22.28 (3.77)	24.00	11	30

White (*n* =28214)										
	Female (*n* =13803)					Male (*n* =14411)				
	n	Mean (*SD*)	Med	Min	Max	*n*	Mean (*SD*)	Med	Min	Max
Pre	4	21.00 (2.94)	20.50	18	25	4	18.75 (3.77)	18.00	15	24
Beg	12	17.58 (3.34)	18.00	12	24	28	18.50 (4.63)	18.00	9	29
Inter	11	19.73 (2.76)	18.00	18	24	17	17.71 (3.06)	18.00	12	24
Adv	55	22.18 (4.40)	24.00	12	30	62	20.37 (3.50)	20.00	12	28
Exit	97	22.16 (3.73)	23.00	9	30	83	20.55 (3.59)	19.00	12	30
Non	13624	22.37 (3.50)	24.00	7	30	14217	20.87 (3.89)	21.00	6	30

Hispanic (*n* = 3506)										
	Female (*n* = 1699)					Male (*n* =1807)				
	n	Mean (*SD*)	Med	Min	Max	*n*	Mean (*SD*)	Med	Min	Max
Pre	47	14.72 (4.76)	15.00	6	26	38	13.87 (4.70)	13.00	6	25
Beg	209	17.55 (3.37)	18.00	8	29	291	15.63 (3.70)	16.00	6	26
Inter	126	18.62 (2.86)	18.00	11	25	167	17.52 (3.71)	18.00	7	30
Adv	302	20.72 (3.36)	20.00	6	30	292	19.28 (3.71)	18.00	7	30
Exit	260	21.25 (3.32)	22.00	9	30	258	19.88 (3.66)	19.00	6	30
Non	755	20.47 (3.74)	20.00	6	30	761	19.57 (4.08)	18.00	6	30

Indian (*n* =523)										
	Female (*n* =286)					Male (*n* =237)				
	n	Mean (*SD*)	Med	Min	Max	*n*	Mean (*SD*)	Med	Min	Max
Pre	1	18.00	18.00	18	18					
Beg	44	17.77 (4.05)	18.00	10	30	54	15.37 (4.41)	16.00	6	26
Inter	16	19.56 (2.90)	18.00	16	26	15	16.07 (3.22)	18.00	12	22
Adv	18	19.00 (4.27)	18.00	6	26	18	19.22 (2.62)	19.00	15	24
Exit	33	20.61 (3.05)	20.00	15	26	20	19.85 (3.63)	20.00	12	26
Non	174	20.20 (3.89)	19.00	9	30	130	18.28 (3.97)	18.00	8	28

Pacific Islander (*n* =442)										
	Female (*n* = 214)					Male (*n* = 228)				
	n	Mean (*SD*)	Med	Min	Max	*n*	Mean (*SD*)	Med	Min	Max
Pre						2	22.00 (2.83)	22.00	20	24
Beg	13	16.77 (4.09)	18.00	6	24	15	17.93 (4.68)	18.00	10	24
Inter	8	20.5 (2.62)	20.00	18	24	11	19.45 (3.56)	18,00	12	24
Adv	25	21.56 (3.31)	22.00	16	30	32	19.41 (4.31)	18.00	10	30
Exit	29	22.55 (2.85)	24.00	17	28	20	21.45 (3.32)	22.00	16	28
Non	139	21.73 (3.62)	23.00	9	30	148	20.82 (3.48)	20.00	14	30

African American (*n* = 373)										
	Female (*n* =194)					Male (*n* =179)				
	n	Mean (*SD*)	Med	Min	Max	*n*	Mean (*SD*)	Med	Min	Max
Pre	2	8.00 (2.83)	8.00	6	10	4	10.75 (2.99)	11.00	7	14
Beg	14	16.43 (5.02)	18.00	9	24	4	15.75 (2.06)	16.00	13	18
Inter	3	17.00 (1.73)	16.00	16	19	4	15.75 (2.06)	16.00	13	18
Adv	5	20.00 (3.74)	18.00	16	24	4	20.75 (3.30)	21.00	17	24
Exit	3	20.67 (4.62)	18.00	18	26	7	17.86 (2.97)	18.00	13	21
Non	167	20.90 (3.31)	21.00	12	30	156	18.92 (3.88)	18.00	9	30

APPENDIX E

2007 DESCRIPTIVE DATA FOR TOTAL SCORE ON THE DWA

Asian (*n* = 614)										
	Female (*n* =295)					Male (*n* =319)				
	n	Mean (*SD*)	Med	Min	Max	*n*	Mean (*SD*)	Med	Min	Max
Pre	5	18.40 (5.13)	18.00	12	24	4	14.50 (7.14)	13.50	7	24
Beg	13	18.23 (5.04)	18.00	6	24	17	17.47 (5.32)	18.00	8	26
Inter	6	21.00 (2.53)	21.50	18	24	7	18.43 (4.08)	18.00	12	24
Adv	38	21.34 (3.74)	23.00	12	27	56	22.21 (3.78)	24.00	13	30
Exit	71	23.94 (2.81)	24.00	18	30	68	23.25 (4.22)	24.00	12	30
Non	162	23.50 (3.56)	24.00	14	30	167	22.09 (3.61)	24.00	12	30

White (*n* =29898)										
	Female (*n* =14621)					Male (*n* =15277)				
	n	Mean (*SD*)	Med	Min	Max	*n*	Mean (*SD*)	Med	Min	Max
Pre	5	15.40 (4.67)	12.00	12	21	7	12.14 (3.48)	12.00	8	18
Beg	15	20.87 (5.04)	20.00	13	30	12	18.00 (4.92)	18.00	8	24
Inter	8	19.38 (3.42)	18.00	14	24	11	18.82 (6.01)	20.00	8	29
Adv	38	20.63 (3.54)	20.00	12	28	45	20.96 (3.91)	21.00	12	26
Exit	60	22.28 (2.97)	23.00	17	30	82	21.57 (3.80)	24.00	10	30
Non	14495	22.46 (3.60)	24.00	6	30	15120	21.01 (4.03)	21.00	6	30

Hispanic (*n* = 4262)										
	Female (*n* = 2022)					Male (*n* =2240)				
	n	Mean (*SD*)	Med	Min	Max	n	Mean (*SD*)	Med	Min	Max
Pre	43	15.00 (5.40)	16.00	6	24	55	12.62 (4.78)	12.00	6	24
Beg	225	17.17 (4.04)	18.00	7	30	354	16.03 (4.00)	17.00	6	28
Inter	158	19.06 (3.90)	18.00	8	30	212	18.11 (4.06)	18.00	6	30
Adv	368	20.75 (3.69)	20,00	12	30	409	19.65 (3.74)	19.00	6	30
Exit	356	21.78 (3.40)	22.00	12	30	285	20.37 (4.21)	20.00	8	30
Non	872	21.25 (3.90)	22.00	6	30	925	19.52 (4.33)	19.00	6	30

Indian (*n* =624)										
	Female (*n* =282)					Male (*n* =342)				
	n	Mean (*SD*)	Med	Min	Max	*n*	Mean (*SD*)	Med	Min	Max
Pre	1	24.00	24.00	24	24	1	24.00	24.00	24	24
Beg	5	21.00 (3.00)	22.00	18		21	17.19 (3.23)	18.00	12	24
Inter	2	21.00 (4.24)	21.00	18	24	6	17.83 (4.40)	18.00	12	24
Adv	8	20.63 (2.97)	19.50	18	24	16	19.00 (4.44)	19.00	9	30
Exit	8	19.88 (4.58)	20.00	14	26	6	22.00 (5.06)	22.00	16	30
Non	256	20.37 (4.39)	20.00	7	30	292	18.38 (4.58)	18,00	6	30

Pacific Islander (*n* =547)										
	Female (*n* = 281)					Male (*n* = 266)				
	n	Mean (*SD*)	Med	Min	Max	*n*	Mean (*SD*)	Med	Min	Max
Pre	2	17.00 (1.41)	17.00	16	18	1	13.00	13.00	13	13
Beg	16	18.00 (3.08)	18.00	13	24	18	18.11 (3.08)	18.00	12	24
Inter	6	19.83 (3.31)	19.50	15	24	6	18.00 (3.63)	18.00	13	24
Adv	39	21.28 (3.16)	21.00	14	26	33	19.88 (2.96)	18.00	14	24
Exit	47	22.15 (4.00)	23.00	14	30	36	21.25 (3.86)	21.00	12	30
Non	171	21.60 (3.44)	21.00	12	30	172	20.28 (3.94)	20.00	10	30

African American (*n* =438)										
	Female (*n* = 203)					Male (*n* = 235)				
	n	Mean (*SD*)	Med	Min	Max	*n*	Mean (*SD*)	Med	Min	Max
Pre	1	16.00	16.00	16	16	4	8.75 (2.22)	8.00	7	12
Beg	5	15.00 (2.83)	14.00	12	18	9	15.33 (4.87)	13.00	10	24
Inter	1	30.00	30.00	30	30	4	20.75 (2.75)	20.50	18	24
Adv	7	20.29 (4.23)	22.00	14	24	2	24.00 (8.49)	24.00	18	30
Exit	7	19.57 (2.23)	19.00	18	24	3	18.67 (2.08)	18.00	17	21
Non	182	21.49 (3.75)	22.00	11	30	213	19.73 (3.89)	19.00	6	30

REFERENCES

Alexander, P.A., & Parsons, J.L. (1991). Confronting the misconceptions of testing and assessment. *Contemporary Education, 62*, 243-249.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, D.C., American Educational Research Association.

Baldwin, D. (2004). A guide to standardized writing assessment. *Educational Leadership, 62* (2), 72-75.

Blanton, L., & Kroll, B. (2002). *ESL composition tales: Reflections on teaching.* Ann Arbor: University of Michigan Press.

Breetvelt, I., van den Bergh, H., & Rijlaarsdam, G. (1994). Relations between writing processes and text quality: When and how? *Cognition and Instruction, 12* (2), 103-123.

Bruning, R., & Horn, C. (2000). Developing motivation to write. *Educational Psychologist, 35* (1), 25-37.

Bryk, K., & Raudenbush, S. W. (1987). Application of hierarchical linear models to assessing change. *Psychological Bulletin, 10*, 147-158.

Bureau of Economic and Business Research. (2003). *The Utah aging initiative: Anticipating the opportunities and challenges of our aging population.* Retrieved September 22, 2009, from http://www.hsdaas.utah.gov/pdf/aging-initiative.PDF

Clyne, M. G. (1987). Cultural differences in organizations of academic texts: English and German. *Journal of Pragmatics, 11* (2), 211-247.

Connor, U. (1996). Contrastive rhetoric: Cross-cultural aspects of second-language writing. In C. Panetta (Ed.), *Contrastive rhetoric revisited and redefined* (p. 4). Mahwah, NJ: Lawrence Erlbaum Associates.

Connor, U. (1996). *Contrastive rhetoric: Cross-cultural aspects of second-language writing*. New York: Cambridge University Press.

Cronbach, L. J., & Meehl, P. E. (1995). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281-303.

Cutler, L., & Graham, S. (2008). Primary grade writing instruction: A national survey. *Journal of Educational Psychology, 100*(4), 907–919.

Elbow, P., & Belanoff, P. (1986). Portfolios as a substitute for proficiency examinations. *College Composition and Communication, 37*, 336-339.

Eggington, E. W. (1987). Written academic discourse in Korean: Implications for effective communication. In U. Connor, & R. Kaplan (Eds.), *Writing across languages: Analysis of L2 text* (p. 153-168). Reading, MA: Addison-Wesley.

Flower, L., & Hayes, J. R. (1980). The dynamics of composing: Making plans and juggling constraints. In L.E. Gregg & E. R. Steinberg (Eds.), *Cognitive processes in writing* (p. 33). Mahwah, NJ: Lawrence Erlbaum Associates.

Freedman, S. W., & Dyson, A. H. (1987). *Research in writing: Past, present, and future* (technical report). Berkeley, CA: Center for the Study of Reading and Writing.

Garcia, G. E., & Pearson, P. D. (1994). Assessment and diversity. *Review of Research in Education, 20*, 337-391

Goertz, M. E., & Duffy, M. C. (2001). Assessment and accountability across the 50 states. Philadelphia, PA: Consortium for Policy Research.

Goodenough, W. H. (1964). Cultural anthropology and linguistics. In C. Panetta (Ed.), *Contrastive rhetoric revisited and redefined* (p. 48). Mahwah, NJ: Lawrence Erlbaum Associates.

Grabe, W., & Kaplan, R. B. (1996). Theory and practice of writing. An applied linguistic perspective. In M. Rose (Ed.), *When a writer can't write* (pp. 1-18). New York: Guilford Press.

Graham, S. (1982). Composition research and practice: A unified approach. *Focus on Exceptional Children, 14*, 1-16.

Graham, S., MacArthur, C., & Fitzgerald, J. (2007). *Best practices in writing instruction*. New York: Guilford Press.

Hacker, D. J. (2010). The effects of delay on text revision. Unpublished manuscript.

Hedeker, D., & Gibbons, R. (1994). A random-effects ordinal regression model for multilevel analysis. *Biometrics, 50*, 993-994.

Hinds, J. (1990). Inductive, deductive, quasi-inductive: Expository writing in Japanese, Korean, Chinese, and Thai. *TESOL Quarterly,* 87-110.

Hinkel, E. (1999). *Objectivity and credibility in L1 and L2 academic writing.* Cambridge, UK: Cambridge University Press.

Hoft, N.L. (1995): *International technical communication. How to export information about high technology*. New York: John Wiley & Sons, Inc.

Hyland, K. (2003). *Second language writing*. Hong Kong: Cambridge UP.

Isernhagen, J., & Kozisek, J. (2000). Improving students' self-perceptions as writers. *Journal of School Improvement, 2,* 3-4.

James, L. A., Abbott, M., & Greenwood, C. (2001). How Adam became a writer: Winning strategies for low-achieving students. *Teaching Exceptional Children, 33* (3), 30-37.

Kaplan, R. (1988). Contrastive rhetoric and second language learning: Notes toward a theory in contrastive rhetoric. In A. C. Purves (Ed.), *Writing across languages and cultures* (pp. 143-154). Beverly Hills, CA: Sage.

Kaplan, R. (2001). Cultural thought patterns in inter-cultural education. In T. Silva & P. Matsuda (Eds.), *Landmark essays on ESL writing* (pp. 11-25). Mahwah, NJ: Lawrence Erlbaum Associates.

Kasper, L. F., & Petrello, B. A. (1997). Responding to ESL student writing: The value of nonjudgmental approach. *Community Review, 14*, 5-12.

Kubota, R. (1992). Contrastive rhetoric of Japanese and English: A critical approach. In U. Connor (Ed.), *Contrastive rhetoric: Cross-cultural aspects of second-language writing* (pp. 28-55). New York: Cambridge University Press.

Leki, I. (1991). Teaching second-language writing: Where we seem to be. *English Teaching Forum,* 19, 8-11.

Leki, I. (1992). Understanding ESL writers: A guide for teachers. In C. Panetta (Ed.), *Contrastive rhetoric revisited and redefined* (p. 5). Mahwah, NJ: Lawrence Erlbaum Associates.

Luke, D. A. (2004). *Multilevel modeling*. Thousand Oaks, CA: Sage.

Lux, P., & Grabe, W. (1991). Multivariate approaches to contrastive rhetoric. *Lenguas Modernas*, *18*, 133-160.

Martin, J. E. (1992). *Towards a theory of text for contrastive rhetoric: An introduction to issues of text for students, practitioners of contrastive rhetoric*. American University Studies: Series XIII, vol. 19. NY.

Matsuda, P. K. (2001). On the origin of contrastive rhetoric. *International Journal of Applied Linguistics*, *11*, 257-260.

Matsuda, P., & Silva, T. (2005). *Second language writing research: Perspectives on the process of knowledge construction*. Mahwah, NJ: Lawrence Erlbaum.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13-105). New York: American Council on Education/Macmillian.

Montano-Harmon, M. (1988) *Discourse features in the compositions of Mexican, English-as-a-second-language, Mexican/American Chicano, and Anglo high school students: Considerations for the formulation of educational policies*. Ph.D. dissertation, University of Southern California.

National Assessment of Educational Progress. (2002). *Nations report card*. Retrieved April 2007 from nces.ed.gov/nationsreportcard/writing/

National Center for Educational Statistics (2002). The *nation's report card: Writing.* National Assessment of Educational Progress. Institute of Education Sciences.

National Center for Educational Statistics (2003). The *nation's report card: Writing.* National Assessment of Educational Progress. Institute of Education Sciences.

National Center for Educational Statistics (2007). The *nation's report card: Writing.* National Assessment of Educational Progress. Institute of Education Sciences.

National Commission on Writing in America's Schools & Colleges. The need for a writing revolution. (2003). *The neglected "R."*: New York: College Entrance Examination Board.

National Writing Project & Nagin, C. (2003). *Because writing matters*. San Francisco: Jossey-Bass.

National Writing Project & Nagin, C. (2006). *Revised and updated edition: Because writing matters*. San Francisco: Jossey-Bass.

Nye, B., Hedges, L. V., & Konstantopoulos, S. (2000). The effects of small classes on

academic achievement: The results of the Tennessee class size experiment. *American Educational Research Journal, 37*, 123-151.

Osterlind, S. J. (2006). *Modern measurement*: *Theory, principles, and applications of mental appraisal*. Upper Saddle River, NJ: Pearson

Pajares, F. (2003). Overcoming learning difficulties. *Reading and writing quarterly,19* (2), 139-58.

Panetta, C. G. (2001). *Contrastive rhetoric revisited and redefined*. Mahwah, NJ: Lawrence Erlbaum Associates.

Pearson Publishing Company, Pearson Technical Report (2008). 2008 Utah direct writing constructed response scoring. Boston, MA: Pearson.

Perin, D., (2007). Best practices in teaching writing to adolescents. In C.A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (p. 243). New York: Guilford.

Piper, D. (1985). Contrastive rhetoric in technical-writing pedagogy at urban institutions. In C. Panetta (Ed.), *Contrastive rhetoric revisited and redefined* (p. 4). Mahwah, NJ: Lawrence Erlbaum Associates.

Purves, A. C. (1988). Writing across languages and cultures. In C. Panetta (Ed.), *Contrastive rhetoric revisited and redefined* (p. 5). Mahwah, NJ: Lawrence Erlbaum Associates.

Ransdell, S., & Levy, C. M. (1996). Working memory constraints on writing quality and fluency. In C. M. Levy & S. Randsdell (Eds.), *The science of writing: Theories, methods, individual differences, and applications* (pp. 93-106). Mahwah, NJ: Lawrence Erlbaum Associates.

Scherff, L., & Piazza, C. (2005). The more things change, the more they stay the same: A survey of high school students' writing experiences. *Research in the Teaching of English, 39* (3), 271-304.

Schoonen, R., van Gelderen, A., de Glopper, K., Hulstijn, J., Simis, A., Snellings, P., & Stevenson, M. (2003). First language and second language writing: The role of linguistic knowledge, speed of processing, and metacognitive knowledge. *Language Learning*. *53*(1), 165-202

Scollon, R. (1991, March). *Eight legs and one elbow. Stance and structure in Chinese compositions*. Paper presented at International Reading Association, Second North American Conference on Adult and Adolescent Literacy, Banff.

Severino, C. (1993). The doodles in context: Qualifying claims of CR. *The Writing Center Journal*, 14 (1), 44-61.

Silva, T. (2001). Toward an understanding of the distinct nature of L2 writing: The ESL research and its implications. In T. Silva & P. K. Matsuda (Eds.), *Landmark essays on ESL writing* (pp. 191-208). Mahwah, NJ: Lawrence Erlbaum Associates.

Spandel, V. (2005). *Creating writers through 6-trait writing assessment and instruction.* Boston: Allyn and Bacon.

Spandel, V., & Stiggins, R. (1990). *Creating writers: Linking assessment and writing instruction.* New York: Longman.

Strickland, D., Bodino, A., Buchan, K., Jones, K., Nelson, A., & Rosen, M. (2001). Teaching writing in a time of reform. *The Elementary School Journal, 101* (4), 385-397.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics, 5th edition*. Boston, MA: Pearson.

Thomas, J. (1993). Countering the "I can't write English" syndrome. *TESOL Journal, 2,* 12-15.

Tolchinsky, L. (2006). The emergence of writing. In C. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 83-95). New York: Guilford Press.

Utah State Office of Education, (2007). *Direct writing assessment: U-PASS*. Retrieved May 2007 from www.schools.utah.gov/assessment/info_dwa.aspx

Weigle, S. (2006). *Assessing writing*. Cambridge, England: Cambridge UP.

Yancey, K. (1999). Looking back as we look forward: Historicizing writing assessment. *Composition and Communication, 50*, 483-503.